Master's thesis
Geography
Geoinformatics

MODELING CROSS-BORDER MOBILITY USING GEOTAGGED TWITTER IN THE GREATER REGION OF LUXEMBOURG

Samuli Massinen

2019

Supervisors:
Tuuli Toivonen
Olle Järv

UNIVERSITY OF HELSINKI
DEPARTMENT OF GEOSCIENCES AND GEOGRAPHY
DIVISION OF GEOGRAPHY

P.O. Box 64 (Gustaf Hällströmin katu 2)
FIN-00014 Helsingin yliopisto

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty	Laitos/Institution– Department
Faculty of Science	Department of Geosciences and Geography

Tekijä/Författare – Author

Samuli Massinen

Työn nimi / Arbetets titel – Title

Modeling Cross-Border Mobility Using Geotagged Twitter in the Greater Region of Luxembourg

Oppiaine /Läroämne – Subject

Geography (geoinformatics)

Työn laji/Arbetets art – Level	Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
Master's thesis	October 2019	80 pp.

Tiivistelmä/Referat – Abstract

The Greater Region of Luxembourg is the largest cross-border labor market in the European Union with the greatest number of cross-border workers in the area. European integration, the Schengen Area, and socio-economical divergences have been the main factors facilitating human cross-border movements in the area and thus the birth and expansion of the borderland community. Despite the freedom of movement, country borders have not been erased and socio-economic divergences have not been levelled. In addition, the spatial extent of the daily movements is not well known. Thus, it is important to study cross-border dynamics and try to separate daily movements from infrequent mobility patterns.

Thus far, cross-border mobility studies have mainly leaned on national registers and census data. These datasets have mostly been too scarce in trying to understand the complexities of cross-border mobility. Many studies have only focused on aggregate-level movement patterns, and the viewpoint of individuals has been missing. Hence, there has been a growing need for individual-level data to be applied in cross-border mobility research.

In this study, a person-based approach is employed using geotagged Twitter Big Data to study spatio-temporal cross-border mobility patterns in the Greater Region of Luxembourg. The aim is to examine how to implement social media in cross-border research as well as how to separate daily cross-border movers from infrequent border crossers and consequently move beyond aggregate-level inspections. Being one of the first studies of its kind, a heuristic programmatic approach is utilized. To the writer's knowledge, social media data sources have not been applied previously to distinguish different cross-border mobility types. All developed scripts in this study are openly available on Digital Geography Lab's GitHub -pages (https://github.com/DigitalGeographyLab/cross-border-mobility-twitter) to promote open science and to introduce new quantitative method tools for cross-border mobility research.

The results show that social media can be implemented in cross-border mobility research, and social media Big Data can provide a relatively good proxy for daily cross-border mobility of people on a regional level. Aggregate-level cross-border mobility patterns and activity location densities correspond closely with previous studies, and outcomes from temporal variation inspections indicate a valid cross-border mover type identification; Twitter users classified as daily cross-border movers seem to be more mobile on weekdays whereas infrequent border crossers on weekends. Daily cross-border mobility patterns also provided new information about the spatial extent of the movements. In addition, heuristic approach resulted in high accuracy in home detection; the "unique weeks" algorithm introduced in this study produced an accuracy of 88.6 % with respect to the ground truth.

Although the results are promising on a regional level, they should be considered in relation to population densities and Twitter use activity; attributes that both vary spatio-temporally and thus can cause bias. Further studies and method development are also needed to draw global conclusions about cross-border mobility; other geographical areas and study settings could result in varied outcomes. In addition, some solutions with data and methods should be considered with a critical stance due to scarcity of valid references. Yet, this study has identified that the coverage of geotagged Twitter data is dependent on data acquisition processes and that Twitter can provide valuable information for cross-border mobility research. In future studies, multi-level data acquisition processes are recommended jointly with person-based approach combining spatio-temporal and content analysis methodologies.

Avainsanat – Nyckelord – Keywords

human mobility, cross-border mobility, social media, big data, the Greater Region of Luxembourg, GIS

Säilytyspaikka – Förvaringställe – Where deposited

HELDA

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty	Laitos/Institution– Department
Matemaattis-luonnontieteellinen tiedekunta	Geotieteiden ja maantieteen laitos

Tekijä/Författare – Author

Samuli Massinen

Työn nimi / Arbetets titel – Title

Valtioiden rajat ylittävän liikkuvuuden mallintaminen geotag-Twitteriä käyttäen Luxemburgin suuralueella

Oppiaine /Läroämne – Subject

Geography (geoinformatics)

Työn laji/Arbetets art – Level	Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
Pro gradu -tutkielma	Lokakuu 2019	80 s.

Tiivistelmä/Referat – Abstract

Luxemburgin suuralue on Euroopan Unionin suurin, valtioiden rajat ylittävä markkina-alue. Euroopan integraatio, Schengen-alue sekä valtioiden sosio-ekonomiset eroavuudet ovat merkittävimpiä tekijöitä, jotka ovat vaikuttaneet valtioiden rajat ylittävän liikkuvuuden kasvuun sekä rajamaayhteisöjen syntyyn ja laajenemiseen. Vapaasta liikkuvuudesta huolimatta valtioiden rajat eivät ole kokonaan hämärtyneet eivätkä sosio-ekonomiset erot valtioiden välillä ole tasoittuneet. Lisäksi päivittäisten liikkeiden spatiaalinen ulottuvuus ei ole tarkalleen tiedossa. On täten tärkeää tutkia valtioiden rajat ylittävää liikkuvuutta sekä dynamiikkaa ja yrittää erottaa toistuvat liikkuvuusmallit epäsäännöllisistä liikkeistä.

Tähän asti valtioiden rajat ylittävän liikkuvuuden tutkimus on nojannut pitkälti kansallisiin rekistereihin sekä väestölaskenta-aineistoihin. Nämä aineistot ovat olleet pääasiassa liian niukkoja, jotta valtioiden rajat ylittävän liikkuvuuden kompleksisuutta voitaisiin kunnolla ymmärtää. Useat aiemmat tutkimukset ovat keskittyneet ainoastaan ylätason liikkuvuusmallien tarkasteluun, ja yksilönäkökulma on jäänyt puuttumaan. Näistä syistä yksilötason aineistoille on ollut kasvavaa tarvetta valtioiden rajat ylittävän tutkimuksen kontekstissa.

Tässä tutkimuksessa tutkitaan valtioiden rajat ylittävän liikkuvuuden spatio-temporaalisia malleja Luxemburgin suuralueella sosiaalisen median Twitter-aineistoa hyödyntäen. Tutkimuksen tarkoituksena on selvittää, kuinka sosiaalisen median aineistoja voidaan jalkauttaa osaksi valtioiden rajat ylittävää liikkuvuustutkimusta. Tavoitteena on erottaa päivittäiset rajanyliliikkujat epäsäännöllisistä matkoista ja täten löytää uutta tietoa, jota kumpuaa ylätasoa syvemmältä. Tutkimus on yksi ensimmäisistä luonteeltaan, minkä johdosta käytetty metodologia on heuristinen. Kirjoittajan tietämyksen mukaan sosiaalisen median tietolähteitä ei olla aiemmin hyödynnetty erilaisten rajat ylittävien liikkuvuusmallien erottamisessa. Kaikki tässä tutkimuksessa kehitetyt ohjelmakoodit ovat avoimesti saatavilla Digital Geography Lab:in GitHub-sivuilta osoitteesta https://github.com/DigitalGeographyLab/cross-border-mobility-twitter.

Tulokset osoittavat, että sosiaalinen media voidaan jalkauttaa osaksi valtioiden rajat ylittävää liikkuvuustutkimusta: regionaalisella tasolla sosiaalisen median aineistoista voidaan louhia päivittäisen liikkuvuuden toteumaa vastaavia malleja. Tässä tutkimuksessa havaitut ylätason liikkuvuusmallit vastaavat aiempien tutkimusten tuloksia, ja tulokset temporaalisesta vaihtelusta indikoivat liikkujamallien luokittelun olevan validi. Päivittäisiksi rajanyliliikkujiksi luokitellut Twitter-käyttäjät vaikuttivat liikkuvan kaikista eniten arkipäivinä, kun taas harvaltaan valtioiden rajat ylittävät yksilöt viikonloppuisin. Päivittäiset rajanyliliikkumiset tarjosivat myös uutta tietoa liikkeiden spatiaalisesta laajuudesta. Lisäksi heuristinen lähestyminen saavutti korkean tarkkuustason käyttäjien kotimaan tunnistuksessa: tässä tutkimuksessa kehitetty ”uniikit viikot” -algoritmi tunnisti käyttäjien kotimaan 88,6 % tarkkuudella.

Vaikka tulokset ovatkin lupaavia regionaalisella tasolla, tulee niitä tarkastella kriittisesti väestötiheyden sekä Twitter-käyttöaktiivisuuden suhteen, jotka kummatkin vaihtelevat spatio-temporaalisesti ja voivat tuottaa vinoumaa. Jatkotutkimuksia sekä metodien kehitystä tarvitaan, jotta johtopäätöksiä voidaan tehdä globaalilla tasolla: muut maantieteelliset alueet ja tutkimusasetelmat voivat tuottaa eriäviä tuloksia. Lisäksi joihinkin tässä tutkimuksessa tehtyihin ratkaisuihin datan ja metodien osalta on syytä suhtautua kriittisesti kunnollisen vertailupohjan puuttumisen vuoksi. Tästä huolimatta tämä tutkimus on onnistunut tunnistamaan, että Twitter-aineiston kattavuus on riippuvaista tiedonhankintaprosessien monikerroksisuudesta, ja että yleisesti Twitter voi tarjota arvokasta tietoa liikkuvuustutkimukseen. Jatkotutkimuksia on suositeltavaa lähestyä yksilöiden näkökulmasta spatio-temporaalisesti täydentäen kokonaisuutta sisältöanalyysi-metodeilla.

Avainsanat – Nyckelord – Keywords

ihmisten liikkuminen, valtioiden rajat ylittävä liikkuminen, sosiaalinen media, big data, Luxemburgin suuralue, GIS

Säilytyspaikka – Förvaringställe – Where deposited

HELDA

Muita tietoja – Övriga uppgifter – Additional information

In Loving Memory of Seppo Massinen (1957–2018), My Father

Table of Contents

ABBREVIATIONS

API	Application Programming Interface
CAP	Complete Automation Probability
GADM	The Database of Global Administrative Areas
GDPR	General Data Protection Regulation
GIS	Geographical Information Systems
GKD	Geographic Knowledge Discovery
GPS	Global Positioning System
GRL	The Greater Region of Luxembourg
HDA	Home Detection Algorithm
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
SQL	Structured Query Language
STATEC	National Institute of Statistics and Economic Studies of the Grand Duchy of Luxembourg

LIST OF FIGURES, TABLES, AND EQUATIONS

1. INTRODUCTION

Mobility has become one of the key aspects in social sciences in the 21st century; people, goods and information seem to be more and more rapidly mobile. This has led to the rise of *a new mobilities paradigm* emphasizing that individuals are increasingly responsible for the movement (Sheller and Urry, 2006). Depending on the movements' spatial and temporal extents as well as contextual and structural factors, a defining feature for mobility can be cross-border character (Kaufmann, 2000; Brunet-Jailly, 2011; Drevon *et al.*, 2016a).

In prevalent geopolitical paradigm, borders are being described as socio-spatial constructs representing human activities in space (Van Houtum, 2005; Brunet-Jailly, 2011; Sohn, 2014). Thus, borders do not represent only hard territorial outlines (Brunet-Jailly, 2011) but express socio-economic differences between communities and their complex interaction in space (Van Houtum, 2005; Sohn, 2014). When global economy shapes policies between nations and a need for multi-level governance emerges, territorial co-operation areas and borderland communities start to emerge (Brunet-Jailly, 2011).

One example of this is the Greater Region on Luxembourg where European integration, the Schengen Area, and socio-economic divergences have stimulated cross-border movements (Carpentier, 2012; Gerber, 2012; Drevon *et al.*, 2016a). Today, the territorial co-operation area is the largest cross-border labor market in the European Union with the greatest number of cross-border workers in the area (The Government of the Grand Duchy of Luxembourg, 2018).

Although cyclic cross-border movements have been steadily increasing e.g. in Europe (Drevon *et al.*, 2016a), in the United States and Mexico (Drevon *et al.*, 2016a; Herzog and Sohn, 2016), as well as in Asia (Drevon *et al.*, 2016a), the socio-economic divergences have not been levelled, and the exact spatial extent of these daily movements is not well known (Carpentier, 2012). Hence, it is vital to investigate re-occurring, *daily cross-border movement patterns* and try to separate them from infrequent cross-border movement patterns.

Most of today's scientific research on local and daily cross-border mobility is focusing on the Greater Region of Luxembourg (e.g. Pierrard, 2008; Carpentier, 2012; Gerber, 2012; Melakessou *et al.*, 2015; Drevon *et al.*, 2016a). However, similar studies have also been undergone e.g. in Finland and Sweden (Paasi and Prokkola, 2008), Kenya (Blanford *et al.*, 2015), as well as in the United States and Mexico (Herzog and Sohn, 2016). To this day, border

studies (e.g. daily border crossings) have mainly been studied using qualitative methods (e.g. Paasi, 1999; Paasi and Prokkola, 2008; Huber and Nowotny, 2011; Gerber, 2012) focusing on psychological aspects of borders as well as borders as a social practice. Quantitative perspectives have also been applied (e.g. Pierrard, 2008; Carpentier, 2012; Blanford *et al.*, 2015; Melakessou *et al.*, 2015; Drevon *et al.*, 2016a) but the approach has been predominantly aggregate-flow-based.

What has been missing in the quantitative cross-border mobility studies listed above is *person-based approach*, although the new mobilities paradigm would suggest the opposite. The previous studies have focused mainly on aggregate-level inspections resulting in too general featured outcomes to properly understand the complexities of cross-border mobility.

The reason for this has been a lack of comprehensive data; national statistics, registers, surveys, and census data used in cross-border mobility research have generally been too scarce and inaccurate, although the coverage and usability have varied geographically (Carpentier, 2012; Gerber, 2012; Blanford *et al.*, 2015; Drevon *et al.*, 2016a). Hence, there has been a growing need for individual-level data to be applied in cross-border mobility research, and subsequently to provide correctives and additional information about the phenomenon (Blanford *et al.*, 2015; Drevon *et al.*, 2016a). According to Blanford *et al.* (2015), georeferenced social media data could be one alternative to provide a relative good proxy.

Literature on human mobility utilizing social media data and *Big Data approach* has been growing in recent years too (e.g. Luo *et al.*, 2016; Manca *et al.*, 2017; Rashidi *et al.*, 2017; Toivonen *et al.*, 2019). However, the cross-border character has mostly been missing; Hawelka *et al.* (2014) and Blanford *et al.* (2015) being the vanguards providing first evidence on the applicability of social media data in cross-border research. However, both studies investigated mobility flows only on a macro-level. This has left methodological deficiencies as well as a distinct need to implement Big Data approach to cross-border mobility research. According to Gerber (2012), a theoretical model does not exist that could explain cross-border mobility.

This study investigates spatio-temporal cross-border mobility patterns in the Greater Region of Luxembourg using geotagged Twitter Big Data. It places in the continuum of person-based and spatio-temporally longitudinal mobility studies (e.g. Luo *et al.*, 2016; Martí *et al.*, 2019) adapting Big Data approach. The aim is to examine how to implement social media in cross-border research as well as how to identify different cross-border mover types. The emphasis

here is in trying to separate daily cross-border movers from other border crossers and as a result move beyond aggregate-level inspections in cross-border mobility. In addition, this study is attempting to develop new quantitative method tools for cross-border mobility research to be used and refined in future studies.

Hence, the research questions are as follows:

1) How can georeferenced social media data be used in cross-border research?
    1.1 How to detect home countries and daily life spaces of people?
    1.2 How to extract cross-border movements?
2) What kind of cross-border mobility patterns can be detected spatio-temporally using Twitter data?
3) How can different cross-border mover types be defined and extracted from Twitter data?

The temporal coverage of Twitter data used reaches from September 2010 to December 2018.

To the writer's knowledge, this study is one of the first attempts to investigate daily cross-border mobility with social media data using person-based approach. This raises conceptual as well as methodological challenges related to examining cross-border mobility patterns and extracting different mover types. In this study, these challenges have been tackled with *heuristic approaches* and sentiment analysis.

2. BACKGROUND

2.1 Human mobility

2.1.1 The concept of mobility

Perhaps the most central concept influencing this work is *human mobility*. To understand complex human socio-economical interactions in time and space, one must first assimilate what causes the people to be on the move, what characterizes the phenomenon, and how human mobility is being classified.

According to Kellerman (2012a), human mobility can be described as "shifting" or "the ability to shift" in the most generic sense but all things considered, it is a multifaceted concept covering both spatial (horizontal) and social (vertical) transitions with or without the support of modern technologies. The concept is not only interested in human displacement over space, but it also covers the context and significance of the movement. Hence, it is not just a branch of transportation geography but also a fundamental notion in social science.

Both Hägerstrand (1992) and Kaufmann (2011) describe the movement of people as an essential part of human being's social composition. People move to survive, socialize, and to relax and to enjoy themselves. Thus, the needs and triggers for human movement are diverse. Kellerman (2012b) generalizes this through *push-and pull effects*:

a) Push effects - people have a basic need for **proximity**, **locomotion**, and a tendency to **curiosity**
b) Pull effects – people have a basic need to meet other **people**, visit new **places**, participate in **events**, and seek new **information**.

These basic socio-spatial needs drive people to be on the move. Social mobility is thus highly linked to spatial mobility; when a social transition occurs, it usually also indicates geographical displacement (and vice versa). These affiliations, however, can be complex due to the usage of telecommunication techniques.

Kellerman (2012a) describes *social mobility* as "status transitions of individuals and social groups along societal strata". *Spatial mobility*, on the other hand, is interested in the geographical displacement but also considers the polysemic nature of the movements; spatial displacement does not reveal what underlies it (Kellerman 2012a cit. Kaufmann 2002). Thus,

spatial mobility is being classified by Kaufmann (2005) as a form of mobility dealing with people who travel within a specified geographical area. The notion assumes that an individual is endowed with a certain "social quality". According to Gerber (2012), these social qualities indicate that:

a) A person can be moved or is on the move
b) A person will be able to move and is willing to move

As a result, the overall mobility concept can be referred to as *socio-spatial mobility* or *geographical mobility*, which is classified through spatial and temporal extents. The classification is called Kaufmann's typology (Kaufmann, 2000):

Table 1. Kaufmann's typology; different geographical mobilities.

Spatial / Temporal	**Movement within a catchment area**	**Movement towards the outside of a catchment area**
Cyclic movement	Daily mobility	Travel
Linear movement	Residential mobility	Migration

The spatial extent is divided into movements within and towards the outside of a *catchment area*, the temporal extent into cyclic and linear movements (Table 1). Gerber (2012) defines catchment area as "the smallest possible area within which inhabitants have access to both facilities and jobs".

This work's context is spatio-temporal movements within a catchment area; the focus is on daily mobilities through cyclic movements. Also, this work is a continuation of *person-based studies*; a spatio-temporal analysis approach originating from the 1970s and the concept of time geography (Hägerstrand, 1970).

2.1.2 Person-based approach

Hägerstrand's (1970) *time-space concept* in time geography represents human mobility from the perspective of individuals in which time and space are always interlinked. Each person's life is represented as a *path* in time and space where certain constraints interacting with one another have a fundamental effect on people's daily movements. These limitations in human mobility include:

a) Capability constraints,
b) Coupling constraints, and
c) Authority constraints

Firstly, capability constraints cover limitations in an individual's biological structure (i.e. a need for rest and eating) and mobility tools available (e.g. modes of transportation). Both have a significant effect on *distance*; how far can an individual reach in a certain time limit (e.g. 24 hours)?

Secondly, coupling constraints are related to activities and spatial boundaries; which activities can a person participate in while being physically present in only one place at a time? Advancements in telecommunication techniques have substantially reduced limitations surrounding these constraints. One can now interact e.g. through smartphones despite not being physically present. Finally, authority constraints are limitations related to physical domains. They can be almost permanent (e.g. state boundaries) or contemporary (e.g. a certain seat in a movie theater).

Even though it has been approximately 50 years since Hägerstrand first introduced time-space concept and thus person-based approach in regional science, the socio-economical web model is still valid today and used in present mobility studies as the basis (e.g. Järv *et al.*, 2014; Luo *et al.*, 2016; Miller, 2017). It is important, however, to acknowledge that Hägerstrand's model is a generalization; individuals' mobility patterns fluctuate spatio-temporally (Järv *et al.*, 2014; Willberg, 2019).

The mobility paradigm has also evolved. In the 21st century, mobility has become one of the key aspects in social sciences; people, goods and information seem to be more and more rapidly mobile, which has led to the rise of *a new mobilities paradigm* (Sheller and Urry, 2006).

According to Sheller and Urry (2006), the new mobilities paradigm (also called *mobility turn*) emphasizes that all places on Earth are connected to one another, at least thinly. The whole concept is being described as a hybrid, complex system where different components are increasingly dependent on one another. People construct a web where **individuals are more and more responsible for the movement**. Places are nodes in the web, but they are not static either. The paradigm sees these "immobile infrastructures" (i.e. airports and gas stations) as enablers of the movement - they are like ships navigating through the web. Thus, it can be argued that the objective of person-based studies today is to analyze and understand mobility

patterns and their differences from the individual perspective (Willberg, 2019).

In the context of this work, one of the key terms describing the mobility turn is also *liquid modernity*. Sheller and Urry (2006) describe it as a phenomenon challenging static standpoints in social sciences where states have traditionally been "containers of communities". Liquid modernity emphasizes that the speed of transition between modern societies has become vital, that modern technologies "enable people" to cross state borders even on a daily basis and eventually live one's daily life within several countries. Transportation and communication are much closer to each other than ever before, which further supports increasing mobilities beyond state borders. On the other hand, this digital and liquid modernity means that people are leaving behind more and more their personal digital footprints.

2.2 Cross-border mobility

2.2.1 The concept of border and borderland communities

Connecting the mobility discussion to human movements across state borders first requires an inspection of state border ontology. Van Houtum (2005) explains that a traditional take on borders in geopolitical studies can be described through the concept of *boundary* - a standpoint prevalent in the 1960s where demarcation (i.e. marking off a boundary or setting a limit, evolution of the phenomenon) and the location of borders were the most central aspects in border studies. Today, however, a more modern viewpoint, especially in cross-border mobility studies (e.g. Paasi and Prokkola, 2008; Carpentier, 2012; Gerber, 2012; Drevon *et al.*, 2016a), can be approached through the notion of *border*.

Borders are being described as socio-spatial constructs representing human activities in space (Van Houtum, 2005; Brunet-Jailly, 2011; Sohn, 2014). Thus, borders do not represent only hard territorial outlines (Brunet-Jailly, 2011) but express socio-economic differences between communities and their complex interaction in space (Van Houtum, 2005; Sohn, 2014). The whole notion is strongly linked to both inclusion and exclusion of people since it's part of socio-spatial, cultural, economic, and political fabrics (Brunet-Jailly, 2011; Sohn, 2014). Borders also have a virtual and impalpable dimension through the usage of portable technologies, virtual transactions and tracking of human and commodity movements (Brunet-Jailly, 2011). Hence, both Van Houtum (2005) and Brunet-Jailly (2011) emphasize that borders have become more indistinct and should be studied through human interaction and *space of flows*, a concept first introduced by Castells (2000).

According to Kellerman (2012a), space of flows consist of three layers:

a) Technologies - electronic transactions
b) Places – nodes & centers, and
c) Humans – leaders making decisions and guiding policies

According to Brunet-Jailly (2011), space of flows originates when the global economy shapes policies between nations and a need for multi-level governance emerges. This has a direct link to the birth of *borderland communities* (e.g. the Greater Region of Luxembourg) - functional entities shaped by **human interaction** as well as **contextual and structural factors.**

In terms of understanding the complex nature of borders, both Van Houtum (2005) and Brunet-Jailly (2011) see that today's scientific approaches should extend over traditional alignments; the viewpoint of space of flows and human interaction on individual-level should not be left out. Thus, it can be argued that the person-based approach is a valid standpoint for cross-border mobility research. Van Houtum (2005) also argues that the notions of boundary and border should overlap to better discuss bordering practices as well as the justification and ethics of borders.

In this work, the concept of boundary is the basis for aggregate-level classifications (i.e. home country detection and assignment) to separate movements that occur across state boundaries. Otherwise, bordering practices are being investigated through the notion of border trying to separate daily cross-border movements from other movements.

2.2.2 Daily cross-border mobilities

Daily cross-border mobilities are repetitive human movements where state boundaries are being crossed. Kaufmann (2000) defines *daily mobility* as a form of geographical mobility expressing cyclic movements within a catchment area (Table 1). The cyclic nature of daily mobility designates that the flows are *reversible*; the movements are two-way constituting reoccurring origin-destination pairs (e.g. home-work). Daily movements are thus frequently performed, expressing routine activities, and also manifesting different forms of human activity (Ramadier *et al.*, 2005; Kellerman, 2012a).

Kellerman (2012a) identifies three spheres for daily mobilities in spatial context;

a) Context environment,

b) Movement, and
c) Spatial extent.

Firstly, the *context-environment* is referring to information society affecting daily movements. It manifests itself e.g. through space of flows, globalization, and networking (i.e. mobility turn). Secondly, *movement* covers directionality and speed on top of circularity; movements have certain durations and usually a spatial destination. Thirdly, *spatial extent* covers the potential of the movement and links between humans and locations (Kellerman, 2012a).

Recent studies on cross-border mobility in Europe (Carpentier, 2012; Gerber, 2012; Drevon *et al.*, 2016a) identify two main reasons behind daily movements across state boundaries; **European integration** and **socio-economical divergences**. European integration has been described as an alleviator of spatial constraints - the free movement of individuals within Schengen Area enables persons to travel without passports and border controls within mutual state boundaries, and common euro currency stimulates inter-regional economic transactions. Socio-economical divergences, on the other hand, are expressed as the main movement-enabling factors (Carpentier, 2012; Gerber, 2012; Drevon *et al.*, 2016a). For instance, relatively higher wages in a neighboring country but low residential expenses in home country actuate human cross-border mobility through cross-border commuting (Carpentier, 2012).

Due to these reasons, borderland communities in Europe are expanding in function and form. This is the case also in other geographical areas, e.g. between the United States and Mexico in North America (Herzog and Sohn, 2016) as well as Hong Kong and Shenzhen in Asia (Drevon *et al.*, 2016a).

It is, however, essential to express that even though socio-spatial circumstances would suggest daily cross-border movements to occur, the phenomenon does not always befall. In other words, cross-border mobility is an ambiguous concept – there are legal, geographical, economic, social and cultural aspects affecting an individual's decision to move (Gerber, 2012; Sohn, 2014). For instance, some people are commuters or doing shopping between two countries; some people are tourists visiting a foreign country.

Thus, different *cross-border mover types* exist, and objectives for the movements are different. One of the focuses of this work lies in **daily cross-border mobilities** in the Greater Region of Luxembourg trying to identify repetitive cyclic flows. One manifestation of this is **cross-border commuting.**

According to Gerber (2012), international workforce flows are one of the main aspects of human mobility where cross-border commuters play a central role. Nevertheless, this still raises a question; why is there a distinct need to separate daily cross-border mobilities from other cross-border movements? In a most fundamental sense, the first law of geography can be invoked; "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). In addition, although cyclic cross-border movements have been steadily increasing in recent years, the socio-economic divergences have not been levelled, and the exact spatial extent of these daily movements is not well known (Carpentier, 2012). Hence, it is vital to investigate re-occurring, daily cross-border movement patterns and try to separate them from infrequent cross-border movement patterns.

Separating daily cross-border movements from infrequent patterns is challenging, however. Methods presented in previous studies (e.g. Carpentier, 2012; Blanford *et al.*, 2015; Drevon *et al.*, 2016a) have been suitable for aggregate-level inspections but too simplified for individual-level. Hence, proper references are difficult to find.

Nonetheless, studies have been conducted in Europe focusing on mathematical thresholds to identify cross-border commuters (Strüver, 2002; Gerber, 2012; Gerber, 2012 cit. Orfeuil, 2000). These studies have identified that the Euclidian distance does not exceed 100 km, and in general, the distance varies from 80–100 km depending on the modes of transportation used. However, in the Greater Region of Luxembourg, the Euclidian distances can be fairly short, even as short as 40 km (Gerber, 2012).

In this study, these thresholds are important factors when comparing results to previous studies.

2.2.3 Previous studies

To this day, cross-border mobility studies have focused on quantitative perspectives on different spatiotemporal levels (Pierrard, 2008; Carpentier, 2012; Blanford *et al.*, 2015; Melakessou *et al.*, 2015; Drevon *et al.*, 2016a). Some researchers have also taken more qualitative standpoints (Paasi, 1999; Paasi and Prokkola, 2008; Huber and Nowotny, 2011; Gerber, 2012; Ralph, 2015) focusing on psychological aspects of borders as well as borders as a social practice.

Studies focusing on daily cross-border mobilities (i.e. cross-border commuting) have primarily been conducted in the Greater Region of Luxembourg (Pierrard, 2008; Carpentier, 2012; Gerber, 2012; Melakessou *et al.*, 2015; Drevon *et al.*, 2016a) but similar studies have also been

undergone in Finland and Sweden (Paasi and Prokkola, 2008), Kenya (Blanford *et al.*, 2015), Ireland (Ralph, 2015), Austria (Wiesböck *et al.*, 2016), the United States and Mexico (Herzog and Sohn, 2016), as well as in Portugal (Pires and Nunes, 2018).

Thus far, the biggest challenge in cross-border mobility studies has been a lack of comprehensive quantitative data; data sources (i.e. national statistics, registers, surveys, and census data) have been scarce, inaccurate or even out of date, although the coverage and usability have varied (Carpentier, 2012; Gerber, 2012; Blanford *et al.*, 2015; Drevon *et al.*, 2016a). As an example, Gerber (2012) points out that interoperability between national statistics and surveys has been challenging, and Drevon *et al.* (2016a) state that information on the duration of activities (i.e. timestamps) have been missing. Geographically speaking, register data have been weak in Asia but in Europe they have been able to provide relatively reliable results although not being universal (Drevon *et al.*, 2016a).

Due to these data shortcomings, cross-border mobility studies have resulted in too general featured outcomes to properly understand the complexities of cross-border mobility. Most of the studies have only focused on investigations on an aggregate level; person-based approach has been missing (Drevon *et al.*, 2016a). This is one of the main shortcomings in cross-border mobility research since the mobility turn is suggesting the opposite (Sheller and Urry, 2006). In addition, Brunet-Jailly (2011) clearly state that there is a need to focus beyond traditional viewpoints in border studies and implement the standpoint of individuals.

These issues have led to conceptual and methodological deficiencies in cross-border mobility research and have emphasized the need to investigate alternative data sources. Especially , there has been a growing need for individual-level data to be applied in cross-border mobility research, and subsequently to provide correctives and additional information about the phenomenon (Blanford *et al.*, 2015; Drevon *et al.*, 2016a). Thus, there is a distinct need to study cross-border mobility using novel data sources (e.g. social media Big Data) from an individual perspective.

2.3 Big Data approach

2.3.1 The concept of Big Data

Historically, the definition of Big Data has been extremely vague. According to Kitchin and McArdle (2016), the term was first used in the mid-1990s referring to handling and analysis of massive datasets. Since then, efforts in trying to explain what Big Data represents in more detail

have been undergone resulting in various descriptions. For instance, Uprichard (2013) approached Big Data's nature through a set of v-words (e.g. versatility, volatility, and vibrancy) whereas Lupton (2015) through p-words (e.g. productive, predictive, and personal). The problem in these descriptions, however, is that they dig only into Big Data etymology and lack conceptual clarity.

A systematic review on Big Data was carried out not until a few years ago by Kitchin and McArdle (2016) to understand what constitutes Big Data. Based on their findings, there are many boundaries and forms of Big Data. In other words, there is no single Big Data profile where every dataset labeled as Big Data could fit; there are different "species" of Big Data. However, there are some inherent characteristics of Big Data that can be listed:

Table 2. Big Data characteristics according to Kitchin and McArdle (2016).

Characteristic	Definition
Volume	Data quantity, covering the number of records, data storage required for each individual record and the full data storage required for all records.
Velocity	The speed to collect, manipulate and publish data. Usually real-time.
Variety	Structured, semi-structured or unstructured.
Exhaustivity	A system is recording all information, not just a sample.
Fine-grained	Covers both resolution and unique indexes.
Relationality	Data includes conjunctive factors that enable coupling with other datasets.
Flexibility	Covers both extensionality (information can be altered) and scalability (the size of the data can change rapidly).

According to Kitchin and McArdle (2016), these characteristics are not all found in every dataset but they are the ones most common. The two most distinctive Big Data features are *velocity* and *exhaustivity* meaning that most of the time Big Data refers to data that is:

a) Easy to manipulate and publish, and
b) Where all information has been recorded real-time or with only a minor delay.

Although not included in Big Data characteristics, Kitchin and McArdle (2016) point out that also *veracity*, *value*, and *variability* are important attributes.

Firstly, veracity refers to the truthfulness of the data; although full of rich information, the datasets are usually also messy and can include errors. Secondly, value indicates that the data can be used in many purposes and contexts. Lastly, variability emphasizes the mutability of the data; the meaning of the dataset is continuously switching depending on the context.

These characteristics can be found in many different datasets although labeling a dataset as Big Data is still somewhat ambiguous today. In general, Big Data datasets are different from traditional datasets (i.e. surveys and administrative data) in terms of methods, sampling, quality, repurposing, and management. Examples of Big Data datasets include mobile phone records, social media, websites, and sensors (Kitchin and McArdle, 2016).

2.3.2 Big Data as a novel data source in mobility and person-based research

Thus far, the main Big Data sources used in mobility research have been mobile phone (e.g. Ahas *et al.*, 2010; Järv *et al.*, 2014) and social media data (e.g. Hawelka *et al.*, 2014; Blanford *et al.*, 2015; Luo *et al.*, 2016; Huang *et al.*, 2017; Manca *et al.*, 2017; Hasnat and Hasan, 2018). In general, literature on human mobility utilizing social media data and Big Data approach has been growing in recent years (e.g. Luo *et al.*, 2016; Manca *et al.*, 2017; Rashidi *et al.*, 2017; Toivonen *et al.*, 2019). However, in terms of cross-border mobility, studies utilizing Big Data have mostly been missing, Hawelka *et al.* (2014) and Blanford *et al.* (2015) largely being the vanguards providing first evidence on the applicability of social media data in cross-border research. Both studies investigated mobility flows only on a macro-level; Hawelka *et al.* (2014) studied mobility patterns globally, whereas Blanford *et al.* (2015) focused on spatio-temporal cross-border flows in the surroundings of Kenya.

One of the reasons for cross-border mobility research scarcity has been the limitation of utilizing mobile phone data. According to Blanford *et al.* (2015), "these data are restricted to the phone providers and coverage may not extend beyond country boundaries unless subscribers have the necessary roaming capabilities enabled." Hence, cross-border researchers have been leaning predominantly on social media data.

In terms of person-based approach, Big Data and broad-scale data, in general, have launched a mobility data revolution (Willberg, 2019); more and more data are being recorded through communication technologies thus offering more possibilities for human mobility research. On top of mobile phone and social media data already mentioned, GPS (Global Positioning System) tracking with sensors has been one of the key sources of information (Tenkanen, 2013).

In this study, geotagged social media data from Twitter is being utilized.

2.3.3 Social media data

Social media can be defined as "web-based services that allow individuals, communities and organizations to collaborate, connect, interact, and build a community by enabling them to create, co-create, modify, share, and engage with user-generated content that is easily accessible" (Toivonen *et al.*, 2019 cit. McCay-Peet and Quan-Haase, 2017). From a mobility research perspective, this user-generated content provides non-continuous traces that can be used in the detection of human movements (Vanhoof *et al.*, 2018).

There are several different social media platforms available, including e.g. Facebook, Instagram, Reddit, Flickr, and Twitter. Different platforms also mean different proprietors with varying standpoints on data availability. Currently, many platforms are not offering data openly to the public. However, Twitter is one of the few well-known platforms that is providing access through Application Programming Interfaces (APIs).

Twitter is a microblogging social media service for sharing short messages (max. 270 characters). In 2018, there were 336 million active Twitter users per month. Currently, Twitter provides both a Search API and Streaming API (Toivonen *et al.*, 2019) for programmatic data acquisition.

The Search API provides a timeline endpoint allowing the collection of latest tweets by a user based on screen name or user id. Currently, the timeline endpoint can return up to 3200 most recent tweets with a *rate limit*. This limitation means that a developer can do only 900 requests to Twitter's server in a 15-minute interval (Twitter, 2019).

The Streaming API provides three streaming levels: standard, gardenhose, and firehose. Standard streaming means that there is a possibility to filter a 1 % sample of all real-time tweets using either specific keywords, user ids or geographic bounding boxes. Gardenhose level offers a 10 % sample coverage and firehose level a full coverage (Poorthuis and Zook, 2017). Neither latter streaming levels are publicly open; the access rights need to be applied separately.

Poorthuis and Zook (2017) present two frameworks mainly used in social media data acquisition through an API:

a) Elaborate data collection framework, and

b) "Adhoc" data collection framework.

Elaborate data collection is designed for general data acquisition; system records all data provided by a platform. "Adhoc" approach, on the other hand, is used in specific purposes; a distinctive schema is built before data collection. A system does not record all data but only those arrays equivalent to the defined schema. This study utilizes the first approach through the Twitter Search API.

There are also different ways to approach social media data analysis. Toivonen *et al.* (2019) identify:

a) Spatio-temporal analysis (i.e. location-based and person-based approach),
b) Content analysis (e.g. computer vision and natural language processing), and
c) Social network analysis based on likes, comments, and followers.

This study is based on spatio-temporal analysis utilizing a person-based approach. To be more exact, posts with location information activated (i.e. geotagged) are being studied. The content is also used in the detection of users' home countries. It is also pivotal to point out that in order to geotag a post, a Twitter user must first activate location services on the account - this is turned off by default (Sloan and Morgan, 2015).

2.3.4 Opportunities and challenges

In social sciences, pros and cons related to Big Data have been increasingly accounted for. New scientific journals have been founded to address the issues (e.g. SAGE Journals), and the considerations surrounding the topics are argued to become more and more pivotal (Housley *et al.*, 2014). Many researchers claim that Big Data can provide multiple opportunities in societal and mobility research, but there are also several challenges involved (e.g. Goodchild, 2013; Blanford *et al.*, 2015; Sloan and Morgan, 2015; Kitchin and McArdle, 2016; Poorthuis and Zook, 2017; Zook *et al.*, 2017; Martí *et al.*, 2019; Toivonen *et al.*, 2019).

Opportunities are mainly been seen stemming from data volume, velocity, exhaustivity, flexibility, value and relationality (Kitchin and McArdle, 2016; Martí *et al.*, 2019; Toivonen *et al.*, 2019) as well as coverage and recording durations and locations of activities more comprehensively than traditional datasets (i.e. surveys and national statistics) (Järv *et al.*, 2014; Blanford *et al.*, 2015). Previously, it has been difficult and laborious to collect a representative sample but now systems behind Big Data collection can record all data virtually real-time. In

addition, the data can be used in several contexts due to versatile characteristics; coupling with other datasets is possible through relationality, and different schemas can be extracted depending on the research setting. In today's research, this process is referred to as Geographic Knowledge Discovery (GKD) (Tenkanen, 2013, 2017).

In relation to cross-border mobility research, Big Data can thus be argued to offer a possibility to improve conceptual and methodological deficiencies identified in previous empirical studies (Carpentier, 2012; Gerber, 2012; Blanford *et al.*, 2015; Drevon *et al.*, 2016a). Theoretically, different mobility types can be studied at the same time on various temporal levels although not too many studies have investigated this aspect. However, Blanford *et al.* (2015) argue that social media could provide a relatively good proxy: "geo-referenced tweets ordered in time by an individual represent semi-continuous movement for that individual and because of the volume of tweets that are often sent, they can provide key insights into human movement patterns". One of the main aspects of this study is to dive into these dynamics.

Challenges, on the other hand, have been seen arising from methods and data accessibility (Poorthuis and Zook, 2017; Toivonen *et al.*, 2019), data structure and variety (Kitchin and McArdle, 2016; Poorthuis and Zook, 2017; Martí *et al.*, 2019) as well as representativeness of the data (Goodchild, 2013; Sloan and Morgan, 2015; Martí *et al.*, 2019).

Usually, due to the vast volume of the data, traditional software are unable to process the data masses. Hence, coding skills are required and researchers must focus more and more on data science, which is challenging traditionally needed skills. Poorthuis and Zook (2017) argue that this poses a threat of undermining theoretical and methodological skills. It is also underlined that Big Data datasets are not always easy or free to access since data owners do not necessarily share the same interests as researchers (Poorthuis and Zook, 2017; Toivonen *et al.*, 2019).

Although Big Data flexibility can be seen offering opportunities in mobility research, the actual structure/variety of the data might cause challenges. Usually, the datasets include a lot of unnecessary information in relation to study context, and thus, the wanted schema can be tricky to be extracted (Kitchin and McArdle, 2016; Martí *et al.*, 2019). Also, since all data is being recorded real time, the content can be cluttered, random, and include errors (Poorthuis and Zook, 2017; Toivonen *et al.*, 2019). These issues can hamper e.g. the estimation and extraction of cross-border movements.

The issue of representativeness and transferability of the data is often brought forward in Big

Data studies. A common argument is that case studies are highly constricted to certain, small geographical areas which complicate conclusions to be made covering other locations on Earth (Goodchild, 2013; Martí *et al.*, 2019). Also, according to Martí *et al.* (2019), there are contradicting understandings on whether Big Data is representing the whole population in an area; some argue that e.g. location-based social media (i.e. geotagged posts) covers human activities adequately once an appropriate sample is being extracted, but some state the opposite. However, common to these studies is a notion that verification of representativeness is a troublesome process.

One major challenge is also ethics and protection of data privacy (Zook *et al.*, 2017; Toivonen *et al.*, 2019). A regulation influencing this study considerably is General Data Protection Regulation (GDPR) – an act regulating the handling of personal data in the European Union. GDRP was designed to give better protection for personal details and is being applied in European Union member states since May 2018 (The Office of the Data Protection Ombudsman, 2018). In terms of requirements and ethics, this means that e.g. details on individuals must be made anonymous and the whole dataset must be guarded against re-identification (Zook *et al.*, 2017).

In this study, I have disassociated all data from specific individuals. In addition, I represent all main results using density mapping to guard against the re-identification of individuals.

3. MATERIAL AND METHODS

3.1 Study area

The study area of this work is centered in the Greater Region of Luxembourg, a territorial co-operation area located in Western and Central Europe. In 2019, this administrative area includes:

- The Grand Duchy of Luxembourg,
- Saarland and Rhineland-Pfalz in Germany,
- Lorraine in France, and
- Wallonia in Belgium (covering Ostbelgien, the German-speaking community in eastern parts of Belgium).

The Government of the Grand Duchy of Luxembourg (2018) describes the Greater Region of Luxembourg as the largest cross-border labor market in the European Union with the greatest number of cross-border workers in the area. In 2016 alone, more than 220 000 commuters crossed a border every day, out of which approximately 170 000 individuals were targeting Luxembourg. The employment rate in 2015 was 70.1 %, unemployment rate 7.9 % (The Government of the Grand Duchy of Luxembourg, 2018).

Figure 1 represents the number of cross-border commuters per administrative area in 2015, and Figure 3 the development of cross-border commuting targeting Luxembourg from 1994 to 2010 as activity densities. In addition, Figure 4 shows the most common spatial extents for daily cross-border movements as activity spaces in 2016.

According to the official statistics of the Greater Region of Luxembourg (STATEC, 2016), the territorial co-operation area had circa 11,5 million inhabitants in 2016. Table 3 represents population statistics per individual territory covering also two additional areas included in the study. The complete study area covers the enumerated administrative boundaries of the Greater Region of Luxembourg as well as the state of Nordrhein-Westfalen in Germany and the administrative region of Champagne-Ardenne in France (Figure 2).

From now on, the complete study area is being referred to as *the Greater Region*.

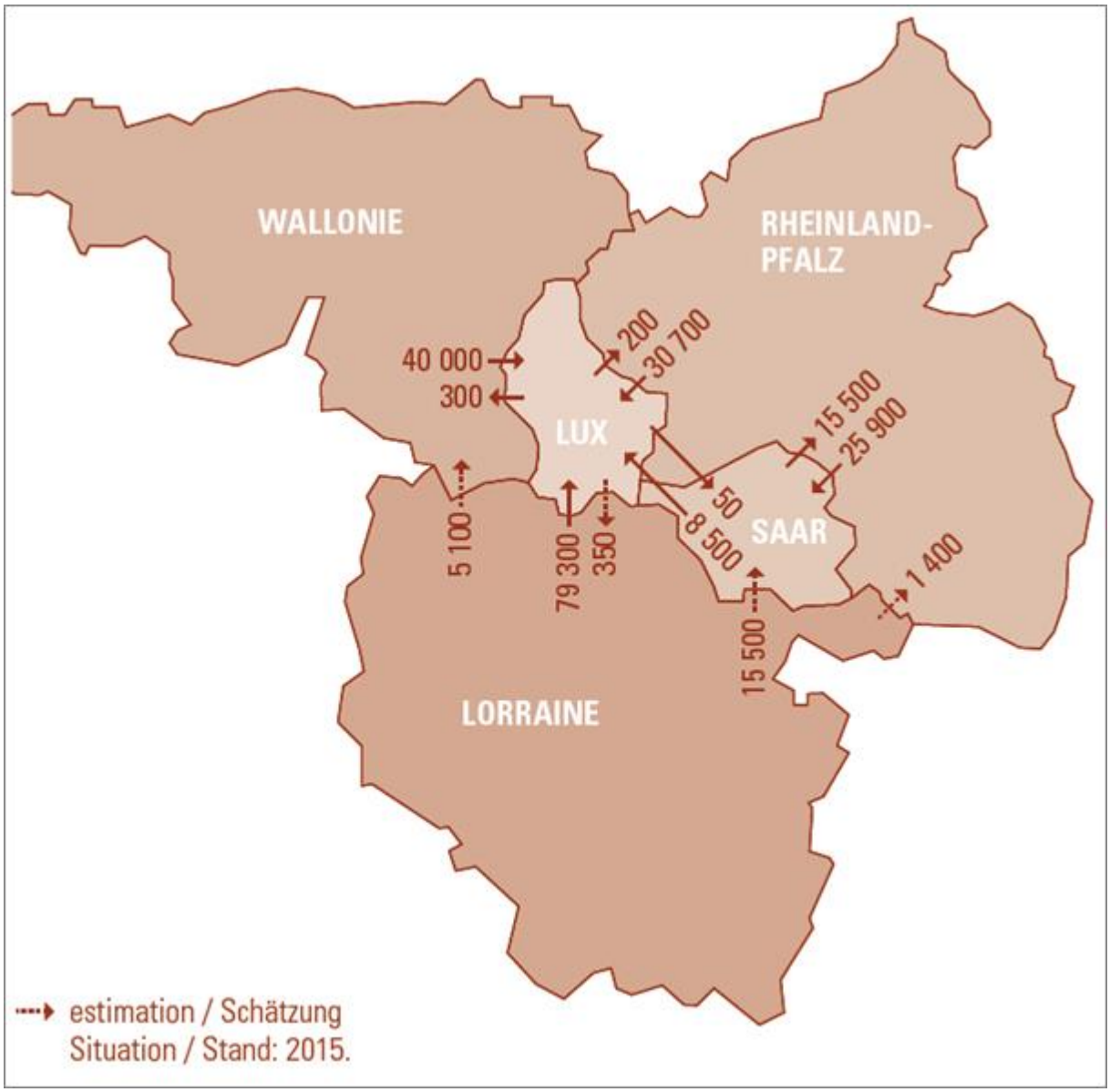


Figure 1. Cross-border commuters in the Greater Region of Luxembourg in 2015 (STATEC, 2016). France has the greatest number of cross-border commuters to Luxembourg, followed by Belgium and Germany.

Table 3. Population in the Greater Region. Information based on 2016 statistics, Nordrhein-Westfalen an exception (2015 statistics). Champagne-Ardenne (European Commission, 2019), Nordrhein-Westfalen (UrbiStat, no date), others *(STATEC, 2016).*

Territory	Inhabitants
Luxembourg	576 249
Saarland	995 597
Rhineland-Pfalz	4 052 803
Lorraine	2 339 019
Wallonia	3 602 216
Nordrhein-Westfalen	17 865 516
Champagne-Ardenne	1 342 363
TOTAL	**30 773 763**

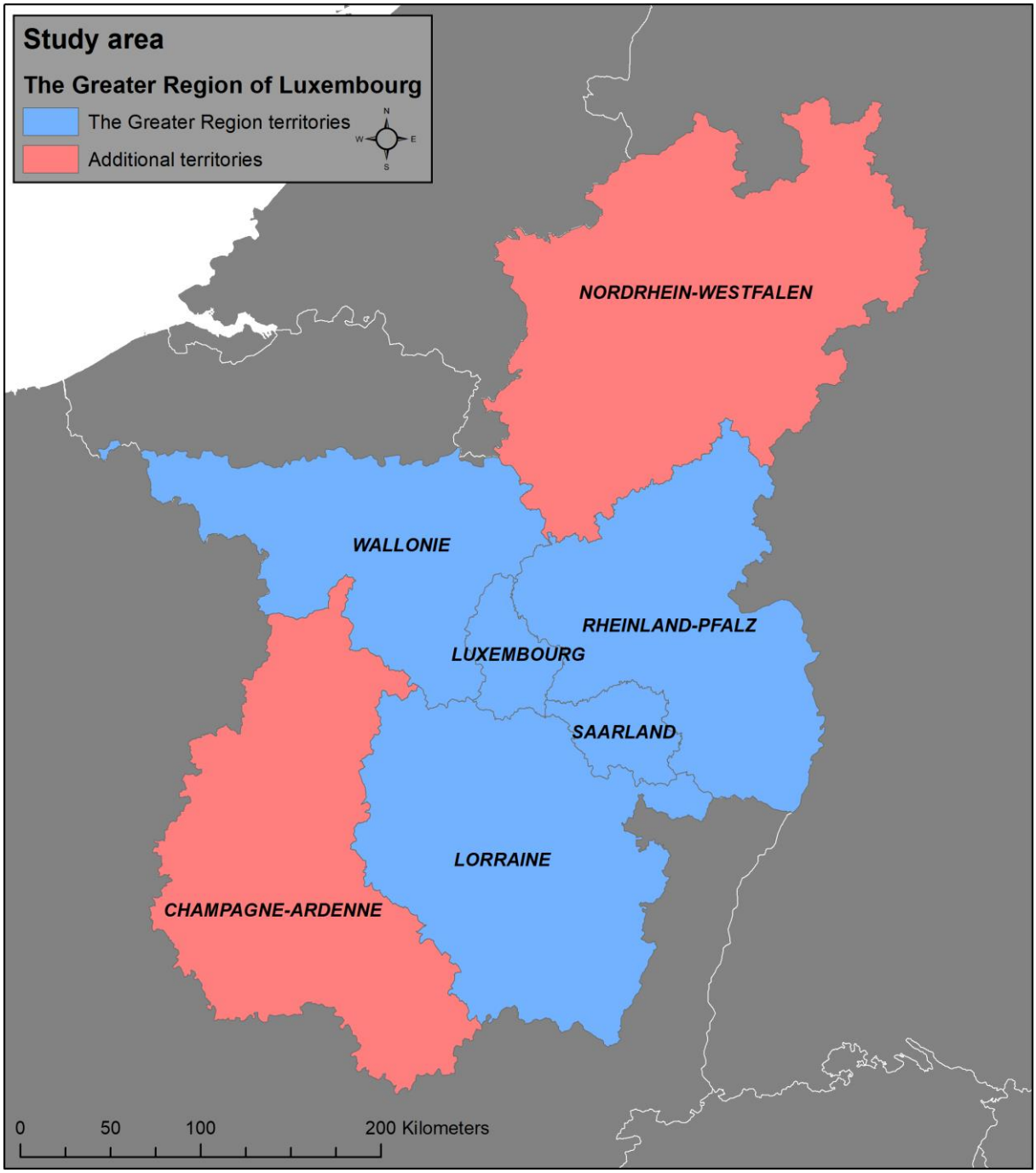


Figure 2. The study area covers the administrative boundaries of the Greater Region of Luxembourg in 2019 as well as two additional territories: Nordrhein-Westfalen in Germany and Champagne-Ardenne in France.

In this study, cross-border movements are being considered between Luxembourg and neighboring nations, not between individual administrative areas. In other words, e.g. Saarland, Rheinland-Pfalz, and Nordhein-Westfalen cross-border movements between Luxembourg both ways are being investigated jointly as Germany-Luxembourg movements.

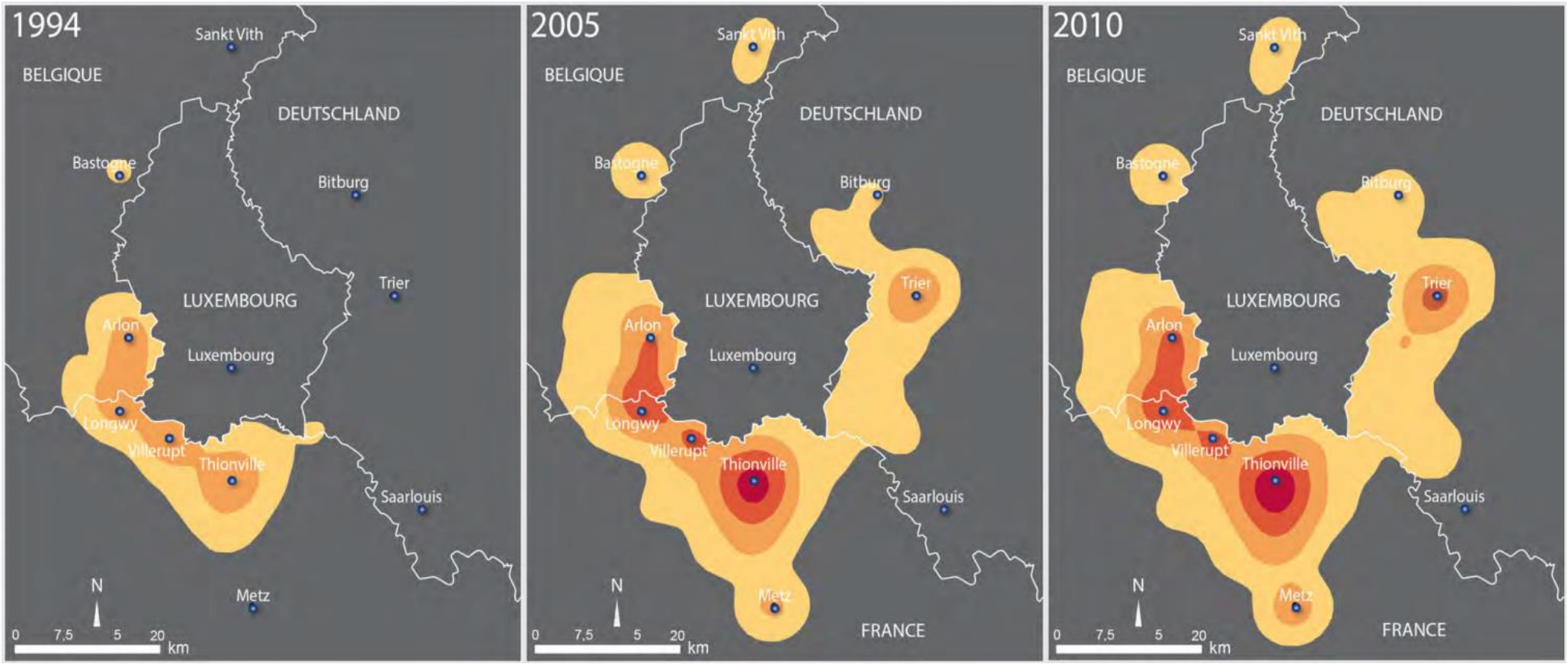


Figure 3. The development of cross-border commuting as activity location densities in the Greater Region of Luxembourg according to Drevon et al. (2016b). One can clearly see that the cross-border activities have been steadily growing.

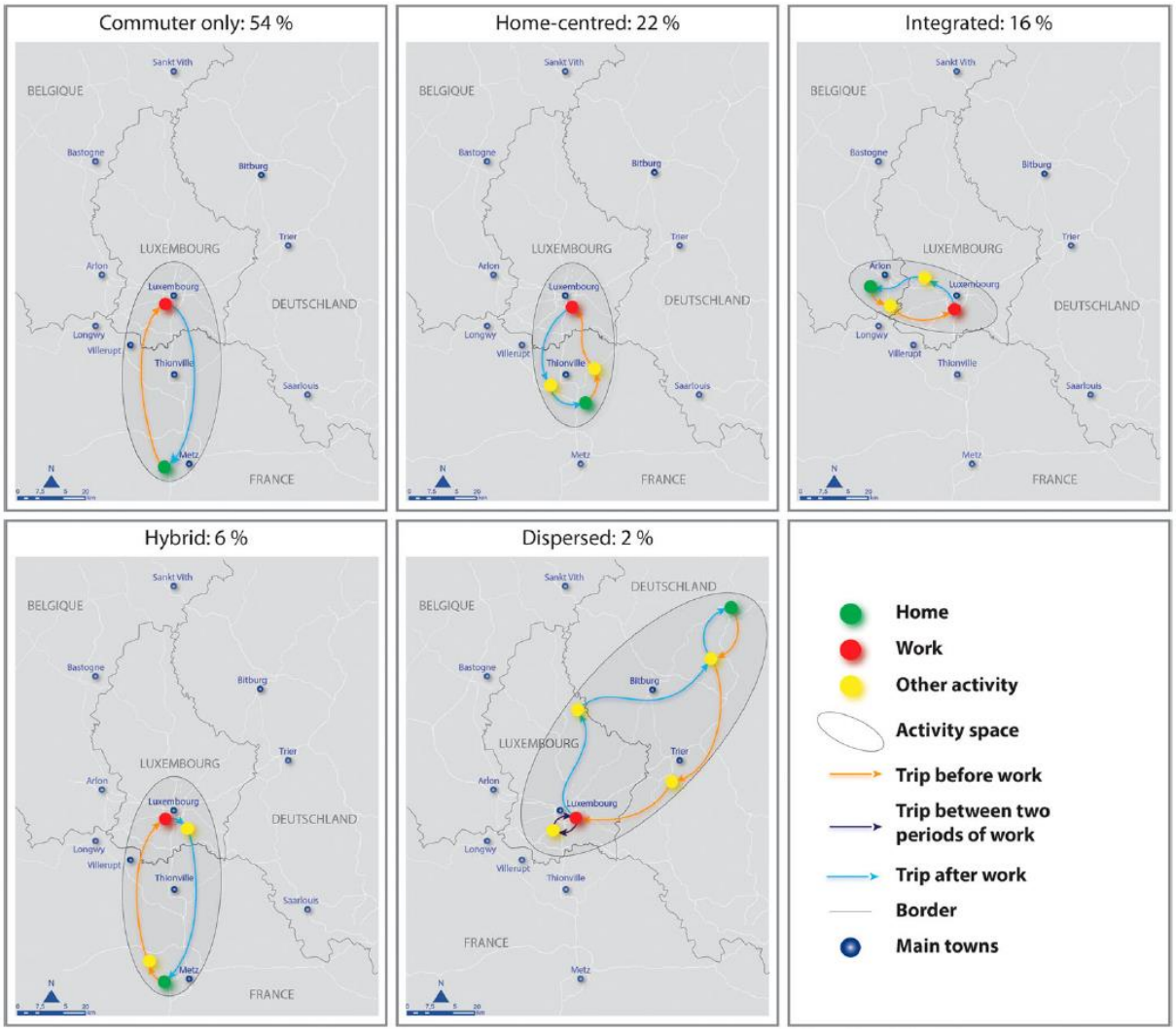


Figure 4. Identified activity spaces and most common spatial extents for different daily cross-border movers in the Greater Region of Luxembourg according to Drevon et al. (2016a).

3.2 Data

3.2.1 Twitter dataset

Digital Geography Lab in the University of Helsinki provided the initial dataset for this study consisting of 1 239 332 publicly available geotagged tweets from 124 994 users in the surroundings of the Greater Region in 2016–2018. This data was collected using Twitter Streaming API from public Twitter user accounts that had posted the tweets openly. This initial dataset was used to identify users who had posted at least once in Luxembourg. To prepare a dataset for this study, tweeting histories of these users were collected using Twitter Search API, covering most recent tweets of all types. The actual Twitter dataset was constructed by filtering out posts without location information, leaving only geotagged tweets in the dataset. In addition, likely bots and found errors were discarded. Thus, the actual Twitter dataset used in the analysis

consisted of 1 022 912 geotagged tweets from 3197 users (Table 4). The temporal coverage of the dataset extended from September 2010 to December 2018. Geotagged tweet counts per each year are presented in Table 5.

3.2.2 Other datasets

In addition to the geotagged Twitter dataset, a global country polygons dataset from Database of Global Administrative Areas (GADM) (2019) was utilized.

The GADM provided spatial data is freely available for academic use including all countries and their sub-divisions. In addition, the dataset's geographic ISO (International Organization for Standardization) codes are in accordance with United Nations Statistics Division (2019) methodology.

3.3 Methods

3.3.1 Study design

I carried out this study using a heuristic programmatic approach to promote open science and to develop new quantitative method tools for future cross-border studies. Data acquisition, processing, and analysis was implemented using Python (version 3.6.7), a programming language designed for writing software in a vast variety of application domains. All scripts used are openly available on Digital Geography Lab's GitHub-pages (https://github.com/DigitalGeographyLab/cross-border-mobility-twitter). The main modules utilized include Tweepy, Pandas, GeoPandas, Shapely, NumPy and Pickle. Visualization was carried out using Esri's ArcMap GIS desktop software.

The complete workflow of the study is presented in Figure 5, and separate analysis phases in their own subsections. As an overall outline, the workflow consisted of three sub-phases: data acquisition and preprocessing, analysis, and creation of outputs. The analysis phase included overall four sub-entities: home detection, detection of cross-border mobility patterns, defining and extracting cross-border mover types, as well as temporal variation inspections. There are three types of visual outputs in this study: maps, charts, and tables.

3.3.2 Protection of personal information

This study is in line with GDPR – an act regulating the handling of personal data in the European Union. For this, I excluded personal information from the user profiles to protect

against the re-identification of individuals. However, the user ids were retained since the ids were pivotal in the data acquisition phase. Yet, I created hashed pseudo ids as part of data collection for each user so that there were zero direct links left for re-identification. Pseudo ids were required for grouping the data by each individual user in the analysis phase.

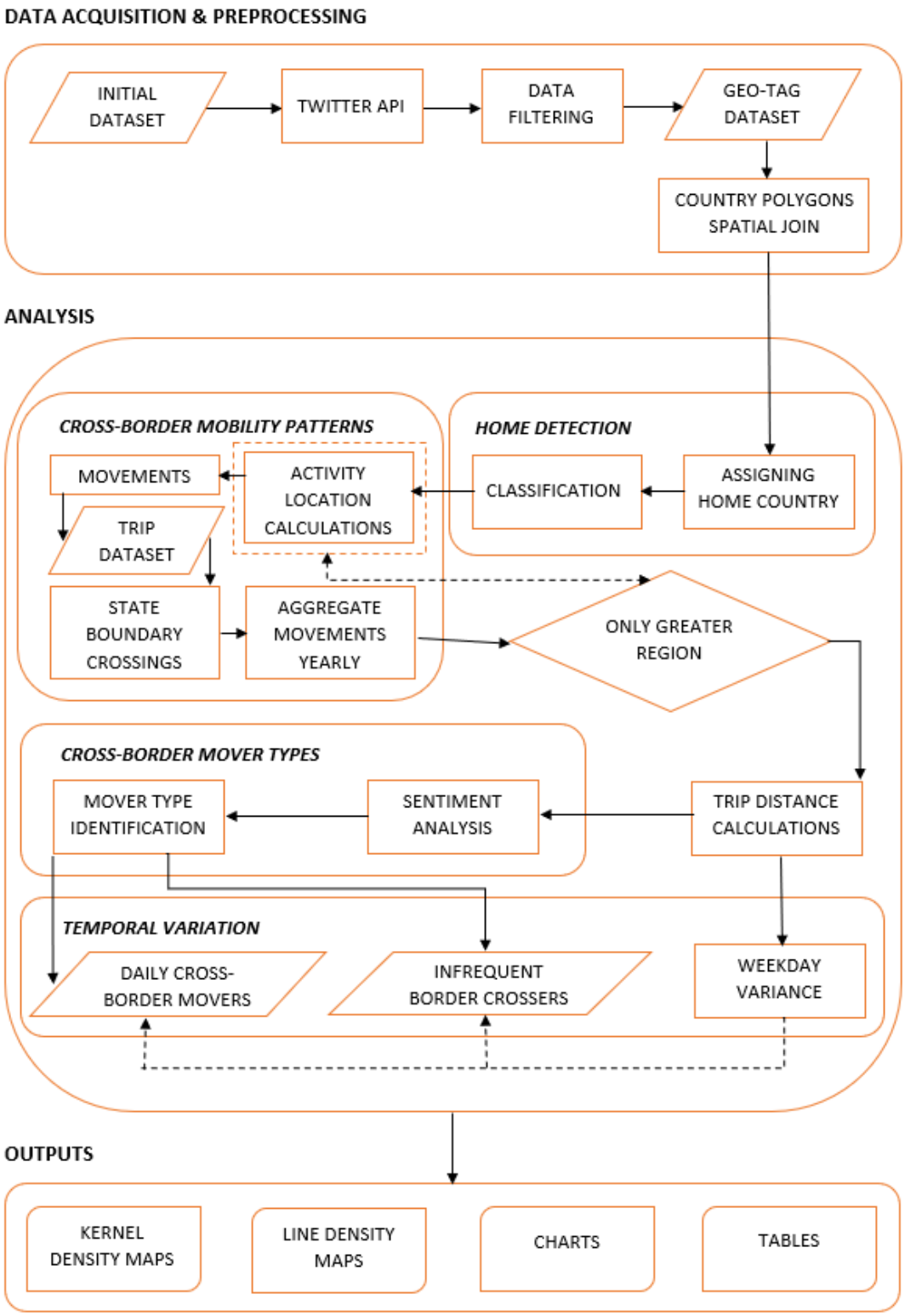


Figure 5. Workflow of the study.

One aspect also considered in this study was guarding against the detection of individual behavior through published results. Re-identification does not necessarily require direct links, latent information can also do harm. In terms of mobility, this sort of predicament can emerge e.g. when presenting trip geometries as LineStrings. If anomalous trips are being clustered spatio-temporally into one catchment area, this might indicate one person's trips and e.g. home-work connectivity. Hence, I present all trips and individual tweeting clusters in this work as density maps (either line or kernel density).

3.3.3 Data acquisition and preprocessing

Digital Geography Lab using Twitter Streaming API on standard streaming level collected the initial dataset. The actual data acquisition for analysis began with the identification of 4020 users who had posted at least once in Luxembourg. The ids of these Twitter users were then saved in a list to gather their tweeting histories using Twitter Search API and Python Tweepy user timeline endpoint. The function at hand returned 3200 most recent tweets per user in JavaScript Object Notation (JSON) format. If an individual did not have 3200 tweets, the function returned all the tweets from the user. An elaborate data collection framework (Poorthuis and Zook, 2017) was used in the data collection meaning that all information provided by the user timeline endpoint was gathered without any extracted schema. A database was not used in data storage. However, all data was packed using Python and Pickle module.

Tweets without location information were then excluded from the dataset leaving only geotagged posts as the basis (~13.5 percentage of the raw Twitter dataset). Some Twitter users are not individual human beings but e.g. automatic bots or advertisement agencies (Hasnat and Hasan, 2018). Hence, bot detection and exclusion was conducted using Botometer, a Python machine learning library developed for identification of automated tweeting activity (Indiana University, 2019). Botometer returns a Complete Automation Probability (CAP) indicating the likelihood of a user being a bot. Based on previous studies (Hasnat and Hasan, 2018; Wojcik *et al.*, 2018), a CAP threshold of 0.40 was selected meaning that users with over 40 % complete automation probability were excluded from the analysis.

The second-to-last phase of data acquisition and preprocessing was the removal of errors. This meant excluding information that lacked unique user information. Altogether 25 704 empty JSON strings were found of this kind. Table 4 describes the data acquisition phase in terms of tweets and unique users.

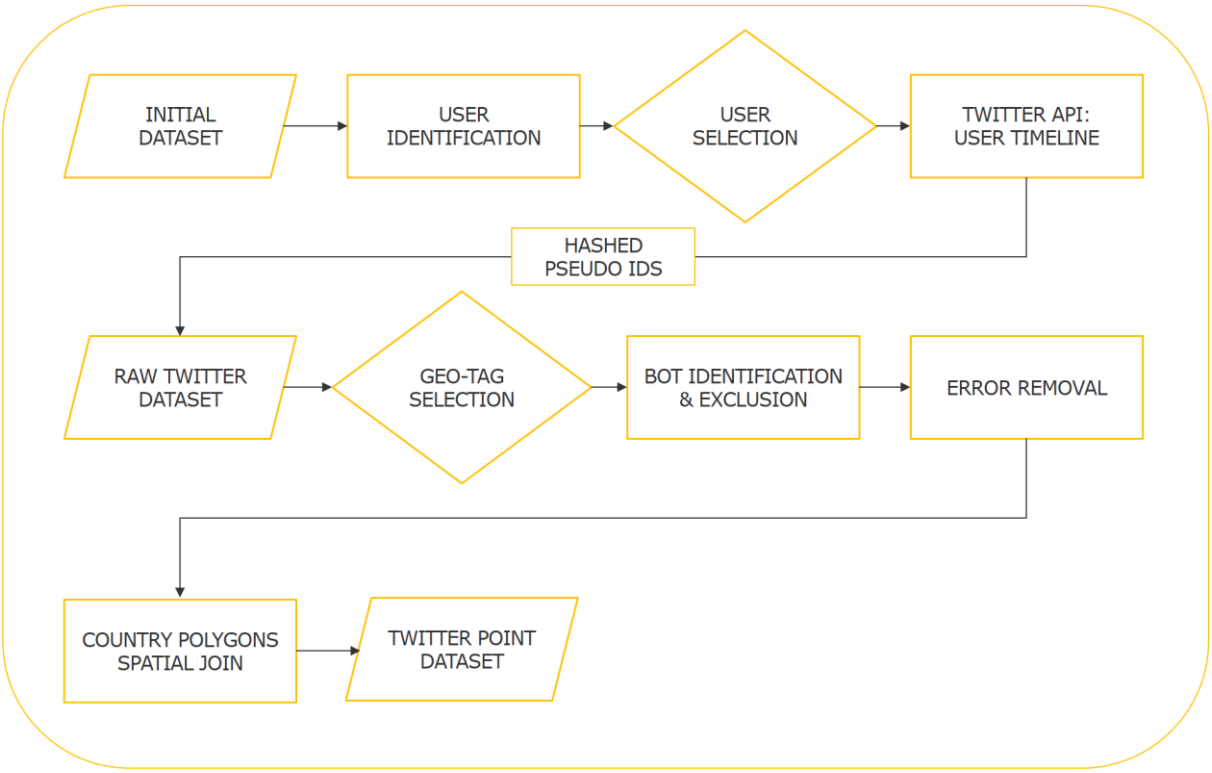


Figure 6. Workflow for data acquisition and preprocessing.

The last phase consisted of spatially joining the Twitter dataset to country polygons shapefile to secure having an accurate information on countries where tweets had been sent.

Table 4. Data acquisition phases and evolution of Twitter dataset in terms of numbers.

Phase	Tweets	Tweets dropped	Users
1) User identification	-	-	4020
2) Twitter API	8 247 548	0	3803
3) Geotag selection	1 110 305	7 137 243	3397
4) Bot exclusion	1 048 616	61 689	3198
5) Error removal	1 022 912	25 704	3197

Table 5. Geotagged tweet counts per each year.

Year	Geotagged Tweets
2018	251 185

2017	257 988
2016	220 236
2015	119 989
2014	89 587
2013	49 138
2012	23 118
2011	11 037
2010	634

3.3.4 Home detection

Home country detection and assignment for Twitter users was based on both user-given information (user profile) and individual tweeting activities (Figure 7). For those individuals who had reported their home location unambiguously and in a country part of the Greater Region (i.e. Belgium, France, Germany or Luxembourg), the home country was assigned directly based on user-given information. Reported content was interpreted to be unambiguous if:

a) A user had announced only one home location (i.e. the information hadn't changed, multiple information wasn't given), and
b) Home location was reported on either country, region or city-level.

Home detection and assignment based on user profile in the Greater Region countries was labeled as the *ground truth* for home detection validation. For the remaining users, the home country was detected and assigned based on tweeting activity using a "unique weeks" Home Detection Algorithm (HDA).

A literature overview of home detection using social media and mobile phone data underlaid the development of the HDA. A total of eight social media (Li *et al.*, 2012; Pontes *et al.*, 2012; McGee *et al.*, 2013; Hawelka *et al.*, 2014; Mahmud *et al.*, 2014; Bojic *et al.*, 2015; Hu *et al.*, 2016; Hasnat and Hasan, 2018) and nine mobile phone articles (Ahas *et al.*, 2010; Frias-Martinez *et al.*, 2010; Frias-Martinez and Virseda, 2012; Phithakkitnukoon *et al.*, 2012; Csáji *et al.*, 2013; Calabrese *et al.*, 2014; Kung *et al.*, 2014; Tizzoni *et al.*, 2014; Vanhoof *et al.*, 2018) were classified in terms of studied home detection methods. After complex decision rules (i.e. machine learning approaches), the most common HDAs used were **spatial groupings**, **time-**

based limitations or **a combination** of these two. Thus, I developed and selected "unique days" and "unique weeks" HDAs for further inspection, out of which the latter one performed better in terms of accuracy. The accuracy calculations were based on the ground truth data.

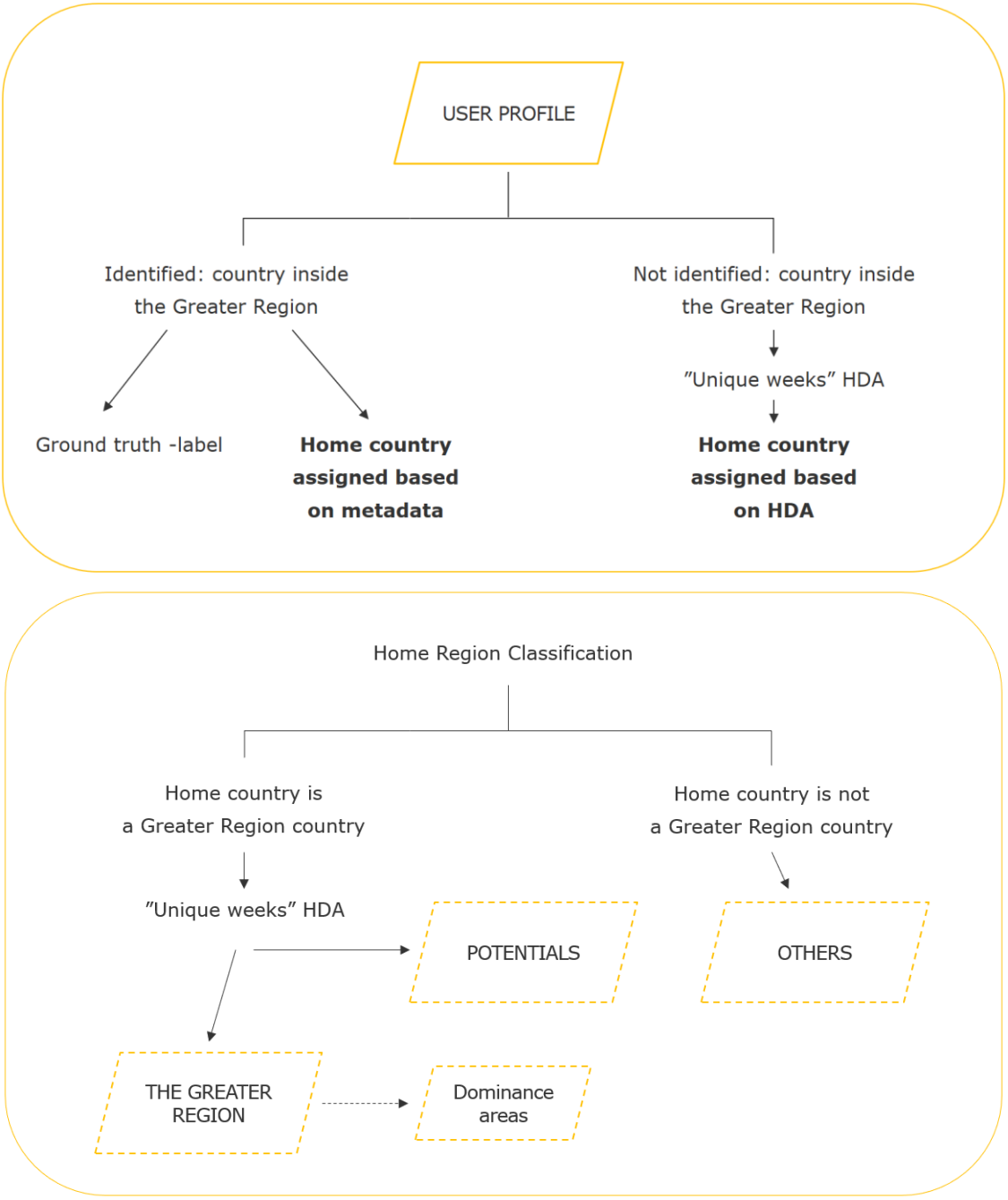


Figure 7. The heuristics for user home country and home region detection.

Once each Twitter user had been assigned to a home country, a three-tier *home region* classification (Figure 8) was implemented:

a) The Greater Region

b) Potentials, and

c) Others

The *Greater Region* class consisted of users whose home location was situated inside the study area. *Potentials*, on the other hand, covered individuals who lived in Belgium, France or Germany but not inside the Greater Region. Lastly, *Others* included all remaining home countries.

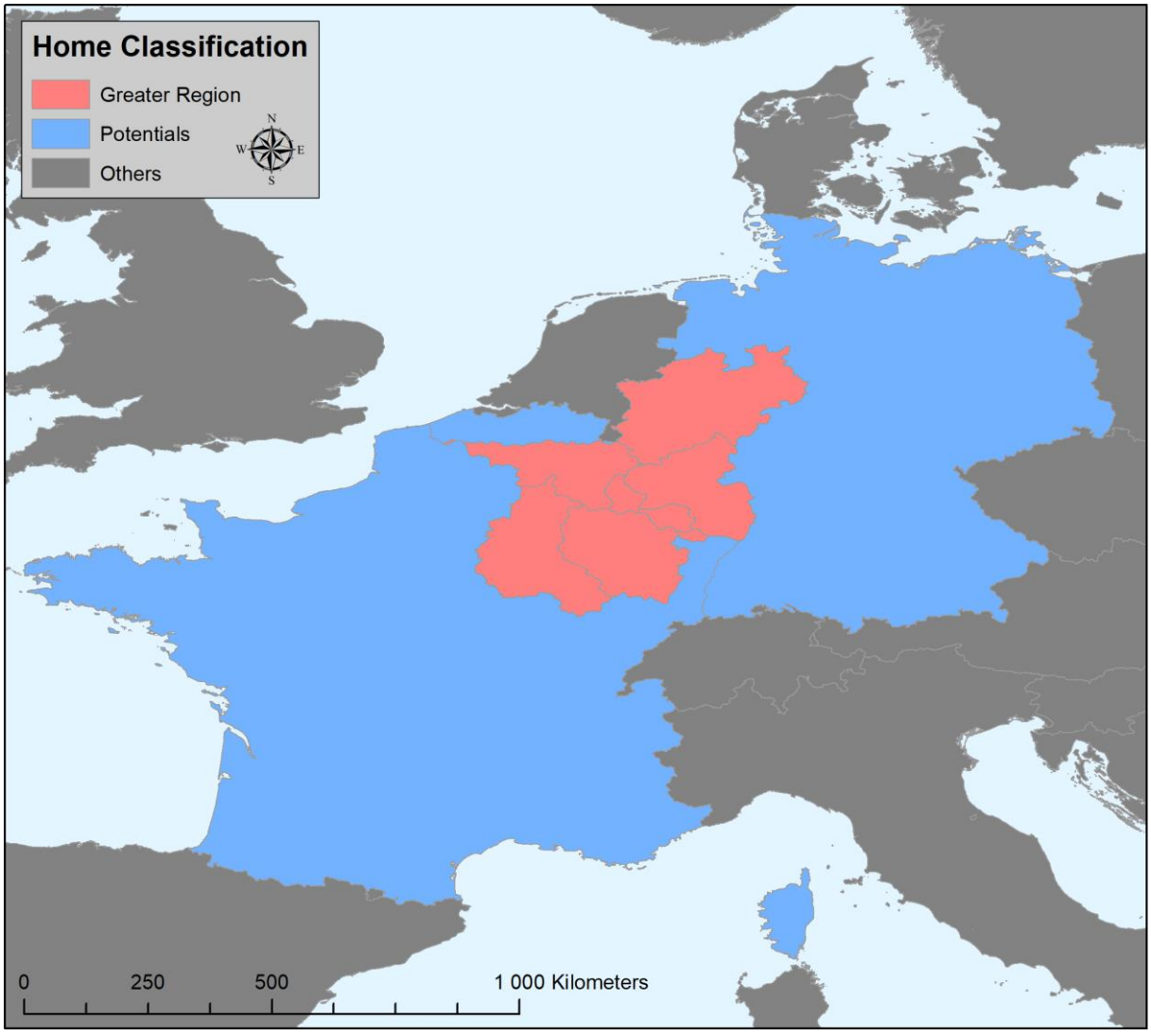


Figure 8. A three-tier home region classification.

The classification was implemented to better understand how distance effects aggregate-level movement patterns – as already mentioned in introduction and background sections, the exact spatial extent of the cross-border movements isn't well known. In addition, separating users living in the Greater Region (i.e. the study area) was pivotal to study daily cross-border movements in relation to Luxembourg. The name Potentials refers to the users' potential to be

cross-border movers in the Greater Region although their home region is not located inside the Greater Region.

Individuals living in Luxembourg were assigned directly to the first class, whereas Belgium, France, and Germany residents had to be divided between the first two classes. The ”unique weeks” HDA developed in this study was implemented again to partition the users. If the HDA detected a user’s activity to be situated inside the Greater Region (e.g. Lorraine in France), the user was attached to the Greater Region class and the activity country section was labeled as the user’s *dominance area*. If two countries inside the Greater Region were tied on top in most tweeting activity, a user was labeled as a *potential cross-border mover*.

3.3.5 Detection of cross-border mobility patterns

Activity location calculations in the Greater Region underlaid the detection of cross-border mobility patterns. The calculations were implemented on both individual and country section levels covering only tweets inside each dominance area. All calculations were implemented using median centroid values.

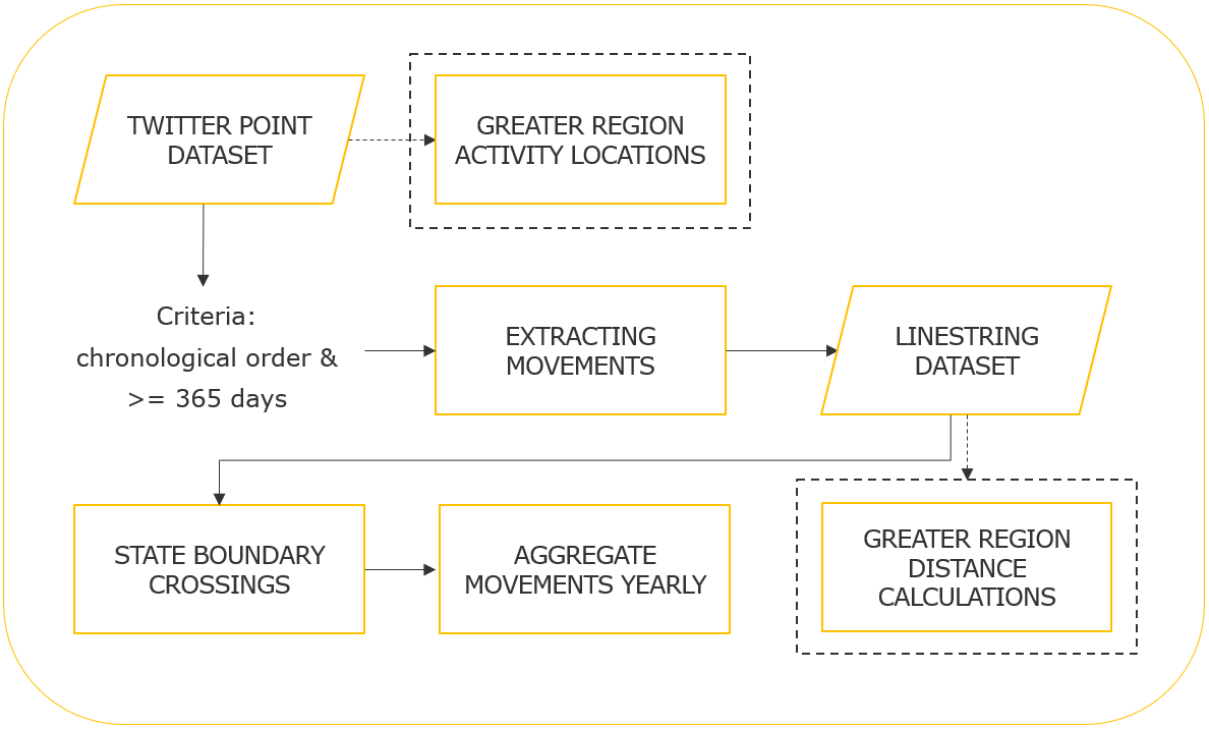


Figure 9. Workflow for detection of cross-border mobility patterns.

The actual detection of cross-border mobility patterns was set up by extracting individual movements from the Twitter dataset (Figure 9). This was based on ordering geotagged tweets

per user in a chronological order. Two consecutive posts were interpreted to represent a trip if under 365 days was passed between the tweets. Both origin and destination countries were stored as attributes to identify cross-border movements between countries.

Extracting individual movements covered also classification of the movements on an aggregate level. The classification was constructed to indicate crossings of state boundaries in the Greater Region:

a) Inside the Greater Region, state boundary crossed ("Inside GRL")
b) Inside the Greater Region, state boundary not crossed ("Inside GRL, no CB)
c) Inbound to or outbound from the Greater Region, the Greater Region administrative area crossed ("Crossing GRL")
d) Outside the Greater Region, state boundary either crossed or not ("Outside GRL"),

Trips distances were also calculated. The calculation was based on Haversine formula ((Equation 1), a mathematical equation used in distance calculations in 3D space (e.g. the great circle distance between two points on a sphere) (Esri, 2017):

$$d = 2r\sin^{-1}\left(\sqrt{\sin^2\left(\frac{\Phi_2-\Phi_1}{2}\right) + \cos(\Phi_1)\cos(\Phi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right)$$

(Equation 1. The Haversine formula.)

Where:

- d = distance
- r = radius of the sphere (in this case Earth, 6371 km)
- Φ_1, Φ_2 latitude of point 1 and point 2
- λ_1, λ_2 longitude of point 1 and point 2

3.3.6 Defining and extracting cross-border mover types

In this study, I developed a heuristic algorithm to extract daily cross-border mobilities for users belonging to the Greater Region home region class. The algorithm processes each Twitter user individually and consists of two inspections:

a) "Inside GRL" trips' share of all trips inside the Greater Region, and
b) Each country section's share of geotagged posts.

The first criterion was passed if “Inside GRL” trips’ share of all trips for a user was >= 20 %. The second criterion, again, was passed if a country section’s share of geotagged posts was >= 20 % and the 20 % threshold was exceeded in at least two country sections. Also, one of these sections had to be a user’s dominance area. This was not self-evident due to already identified possible cross-border movers who didn’t have a dominance area label.

If both criteria were satisfied, a user was labeled as a *daily cross-border mover*. Else, a user was given an *infrequent border crosser* label.

I set up the formulation of the algorithm with a sentiment analysis. Figure 10 represents the relative share of “Inside GRL” trips out of all movements inside the Greater Region, and Figure 11 a country section’s share of geotagged posts where each Greater Region user had posted most of the tweets. For most of the users, this meant the assigned dominance area.

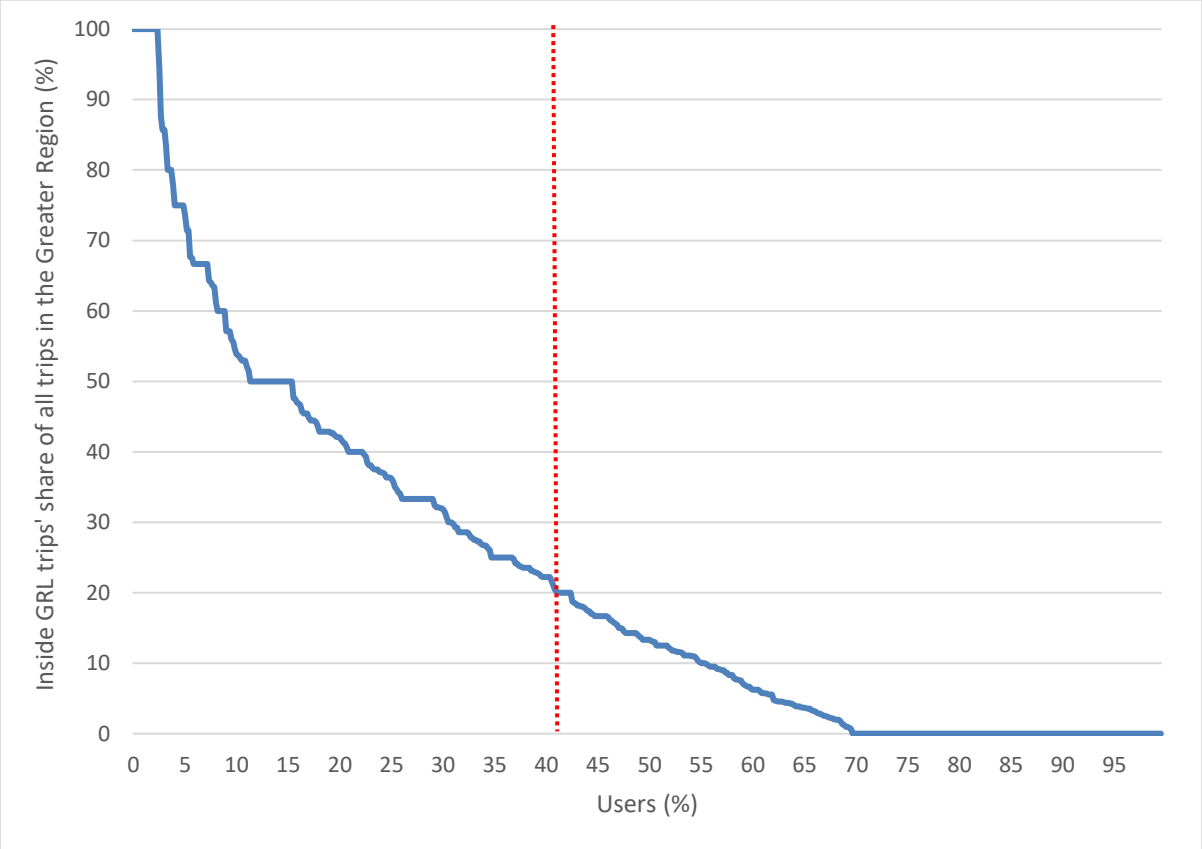


Figure 10. The relative share of state boundary crossings inside the Greater Region for users assigned to the Greater Region home region class.

In attempt to find the smallest common denominator for the two inspections, I detected a 20 % threshold. With “Inside GRL”, it covered approximately 41 % of the users and with country sections’ share, circa 37 %. In other words, selecting 20 % as the threshold for both inspections

covered approximately the same relative share of users. In Figure 11 the 20 % threshold is presented inversely; 80 % is the maximum value to potentially have a 20 % representation in two country sections.

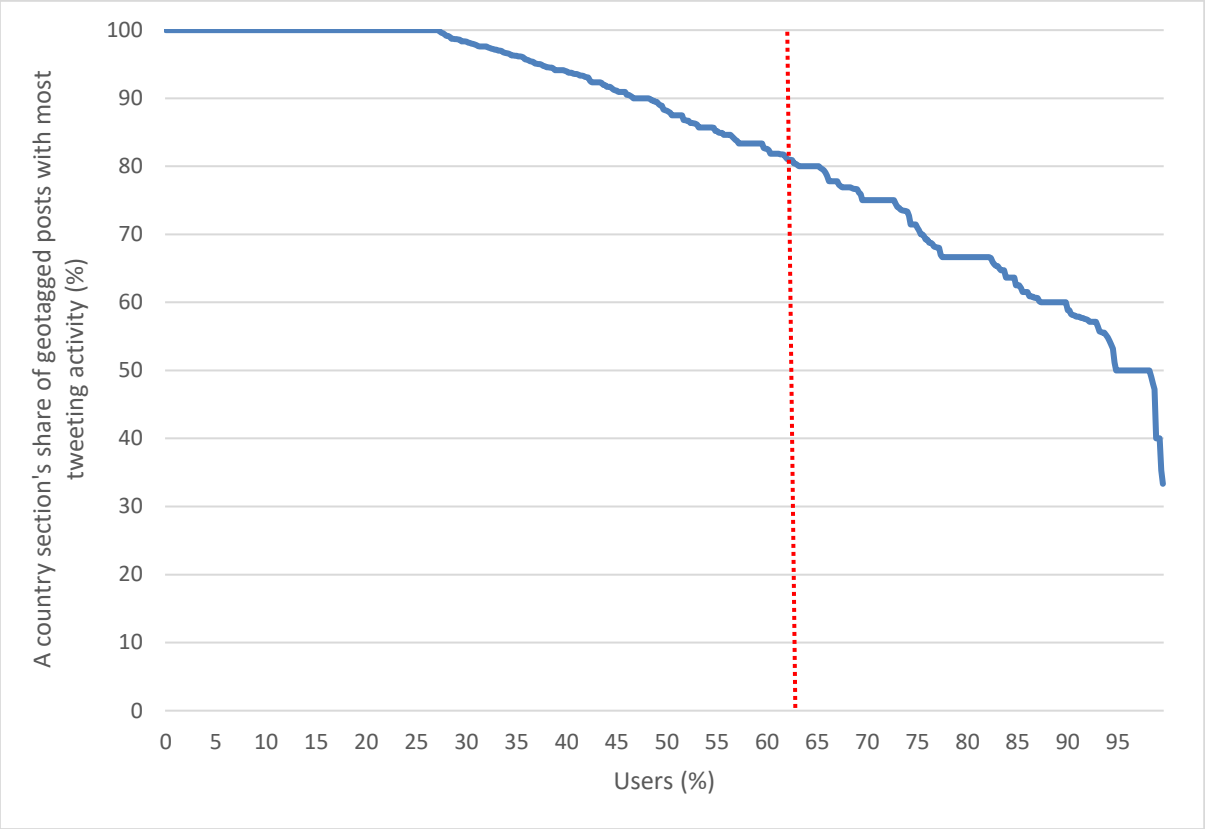


Figure 11. A country section's share of geotagged posts where each Greater Region user had posted most of the tweets. For most of the users, this meant the dominance area.

3.3.7 Temporal variation

An inspection of cross-border movement's temporal variation was conducted for both cross-border mover types. The calculations summed up both origin and destination weekdays from "Inside GRL" trips made both ways between each dominance area and Luxembourg. An average count of cross-border movements was calculated based on all weekdays. Finally, variances were calculated for each weekday in relation to the average value.

4. RESULTS

4.1 Home location detection

For 1052 users out of 3197 (32.9 %), a home country was detected in Belgium, France, Germany or Luxembourg based on user-given information. For the remaining 2145 users (67.1 %), the home country was assigned based on the "unique weeks" HDA developed in this study.

Based on the findings from the literature overview, a method comparison between "unique days" and "unique weeks" was conducted. With respect to the ground truth, "unique weeks" (88.6 %) returned a slightly better accuracy than "unique days" (87.1 %).

In terms of home region classification, 733 users were assigned to the Greater Region class, 819 to Potentials, and 1645 to Others. Table 6 represents the results of the dominance area detection and assignment inside the Greater Region class.

Table 6. Dominance areas inside the Greater Region. User counts as well as AVG and MDN descriptive statistics for geotagged tweets per user.

Dominance area	Users	Average geotagged Tweets per user	Median geotagged Tweets per user
Luxembourg	472	112	22
France	161	98	22
Germany	43	167	32
Belgium	25	34	80
Potential cross-border mover	*32*	*191*	*7,5*
TOTAL	733	111	23

Belgium had the highest median and the lowest average value for geotagged tweets per user. In addition, 32 potential cross-border movers were identified. However, as can be seen from the median values, many users had only few geotagged tweets posted. Only one user with a potential cross-border mover label had several geotagged posts, which had a considerable effect on the relatively high average value (191). The user in question was identified as a daily cross-border mover between Luxembourg and France and was eventually given France as a

dominance area label. Other users labeled as potential cross-border movers were excluded from the later analysis resulting in 702 individuals belonging to the Greater Region class.

4.2 Aggregate-level cross-border mobility patterns

Based on activity location densities, tweeting activity of people is distinctly clustered in the city of Luxembourg when considering all geotagged posts in the Greater Region (year coverage 2010–2018). In addition, the dominance of France is prominent (Figure 12). These findings are in line with the results from dominance area detection and assignment; most of the Twitter users were found having greatest activity in either Luxembourg or France (Table 6).

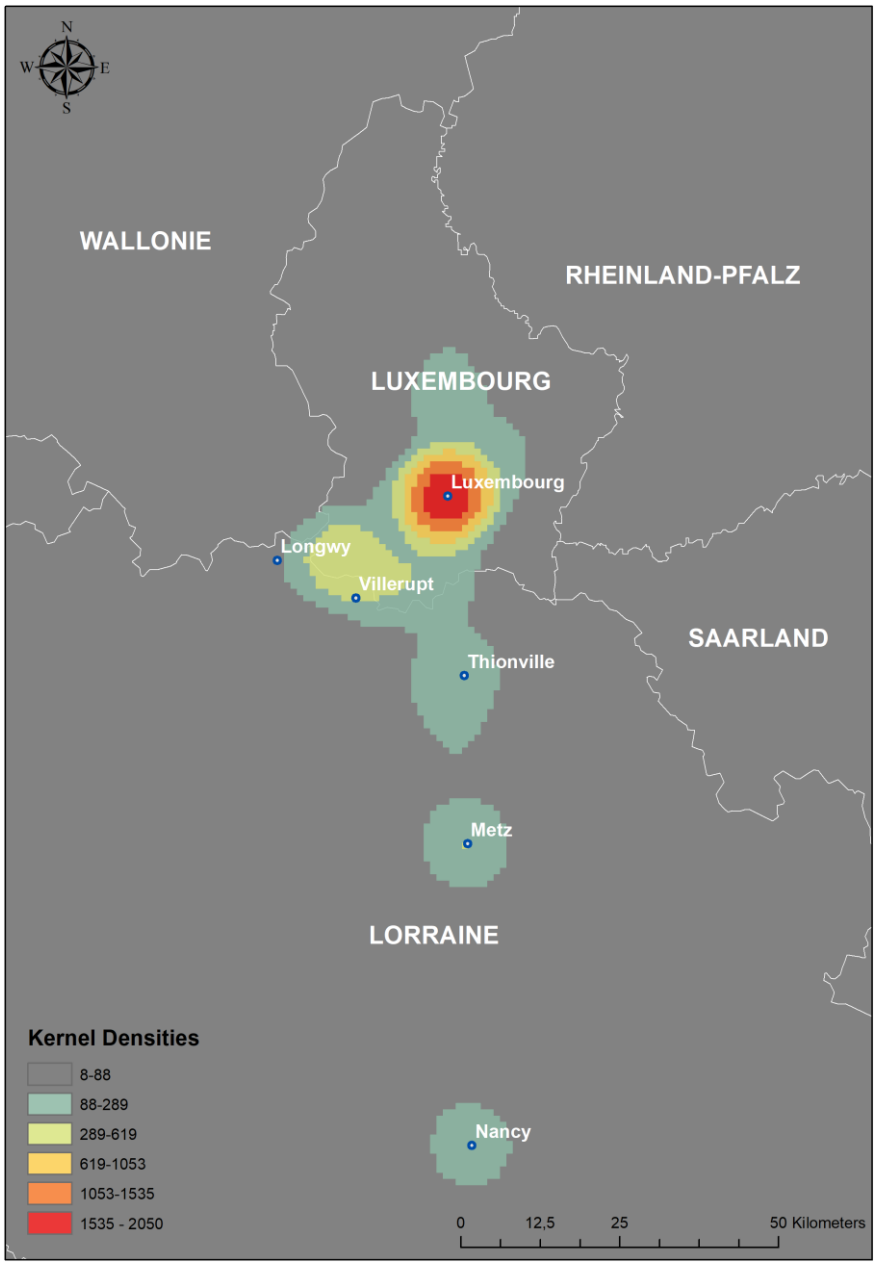


Figure 12. Activity location densities based on user median centroids. The main activity is clustered in the heart of Luxembourg. Kernel density cell size 1000 m^2, 10 000 m search radius.

Overall, 816 033 trips were detected from the used dataset. Figure 13 and Table 7 represent the nature of these movements in relation to the Greater Region state boundaries.

A distinct outcome is that users belonging to the Greater Region class are mainly crossing a state boundary inside the Greater Region. The further away a user is living in relation to the Greater Region, the more common it is that the movements and border crossings are happening outside the administrative area (Figure 13).

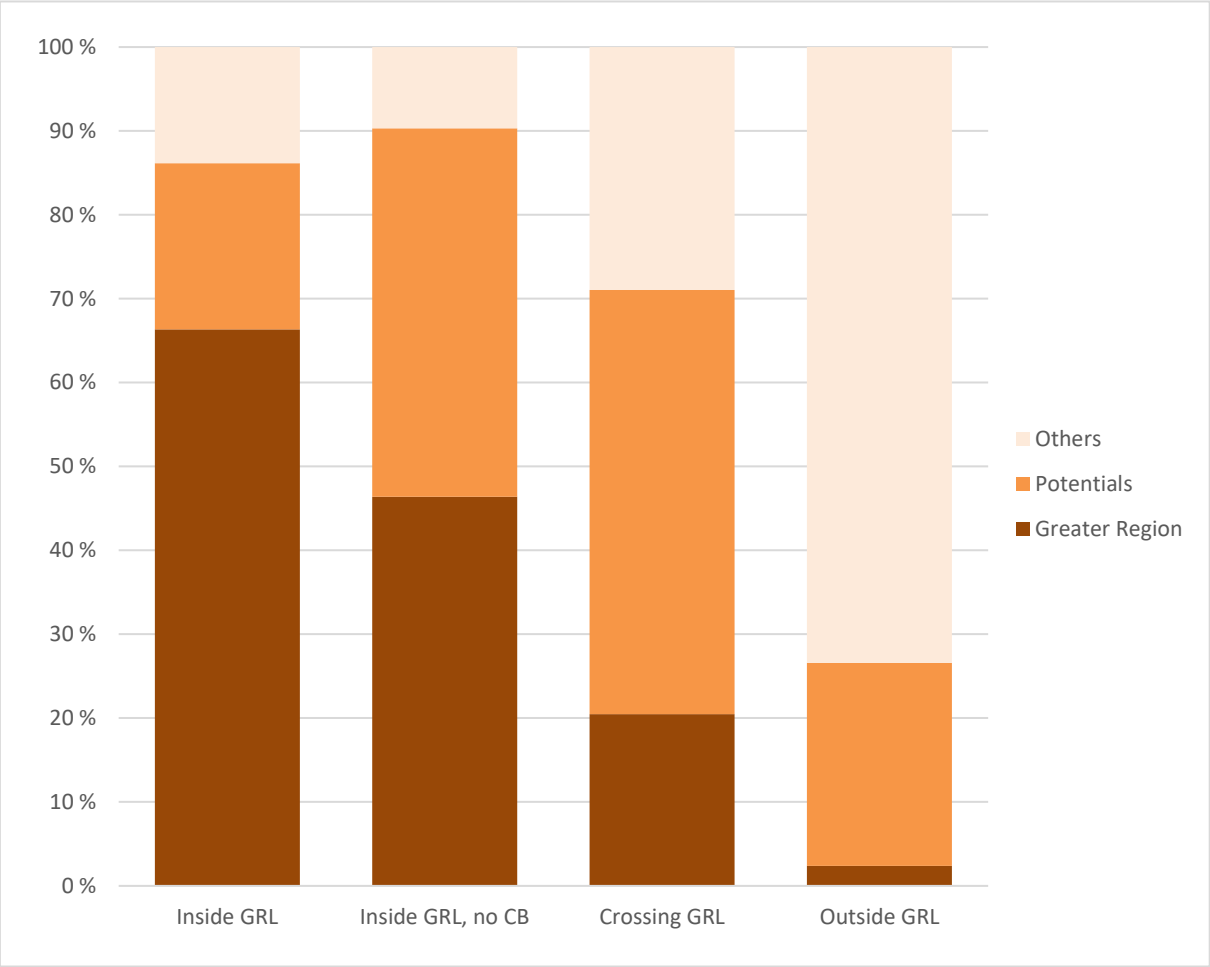


Figure 13. Movement types in relation to state boundaries and the Greater Region. Year coverage 2010–2018.

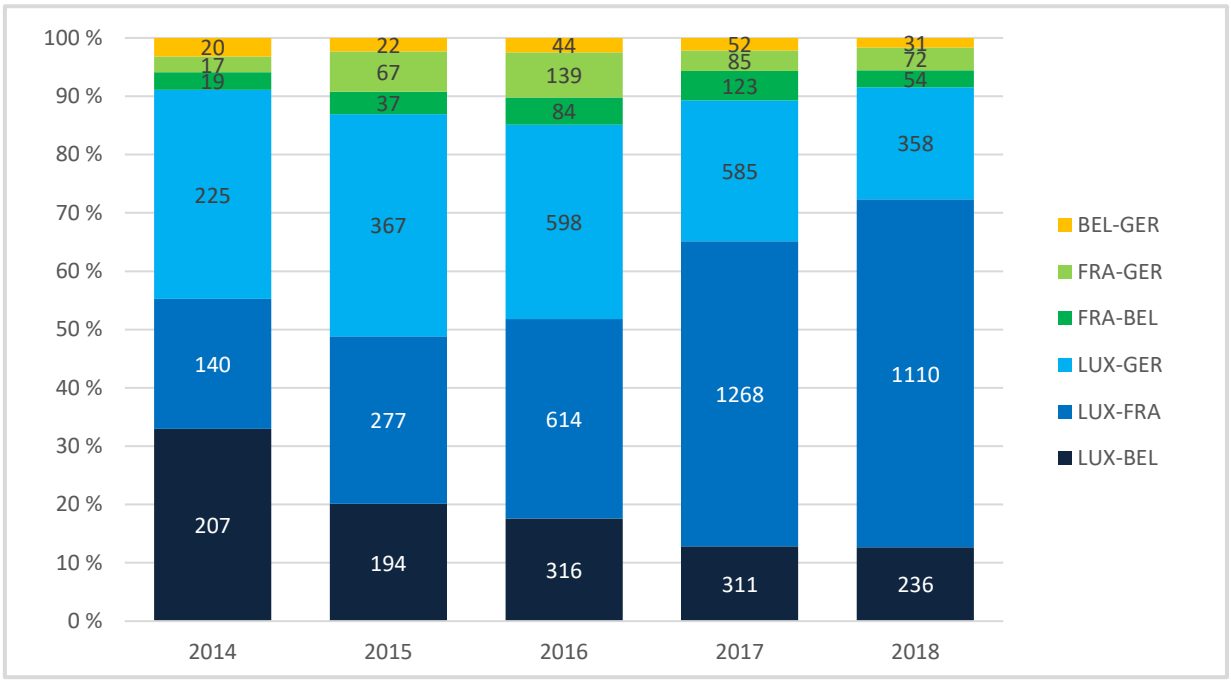
For the Greater Region users, the most common movement is a trip inside the area without a crossing of a state boundary. For Potentials and Others, most of the movements are happening outside the Greater Region. In terms of crossing the Greater Region, the number of trips is relatively minor for all home region groups. The Greater Region users have the largest relative share of 13.5 % while Potentials have 8.8 % and Others only 2.0 % (Table 7).

Table 7. Extracted mobilities and the relative shares of movement types for different home region groups. Year coverage 2010–2018.

	Greater Region		Potentials		Others		TOTAL	
Inside GRL	5776	(10.0 %)	1725	(0.8 %)	1207	(0.2 %)	8708	(1.1 %)
Inside GRL, no CB	26 934	(46.8 %)	25 525	(11.7 %)	5625	(1.0 %)	58 084	(7.1 %)
Crossing GRL	7747	(13.5 %)	19 133	(8.8 %)	10 966	(2.0 %)	37 846	(4.6 %)
Outside GRL	17 059	(29.7 %)	171 940	(78.8 %)	522 396	(96.7 %)	711 395	(87.2 %)
TOTAL	57 516	(100 %)	218 323	(100 %)	540 194	(100 %)	816 033	(100 %)

The following figures (Figure 14 - Figure 17) represent aggregate cross-border movements yearly from 2014 to 2018 based on three-tier home region classification (for reasons behind dropping 2010–2013, see discussion section 5.1.2 Twitter API). The movements cover crossings of the state boundaries between the Greater Region countries, not only between the Greater Region country sections.

THE GREATER REGION

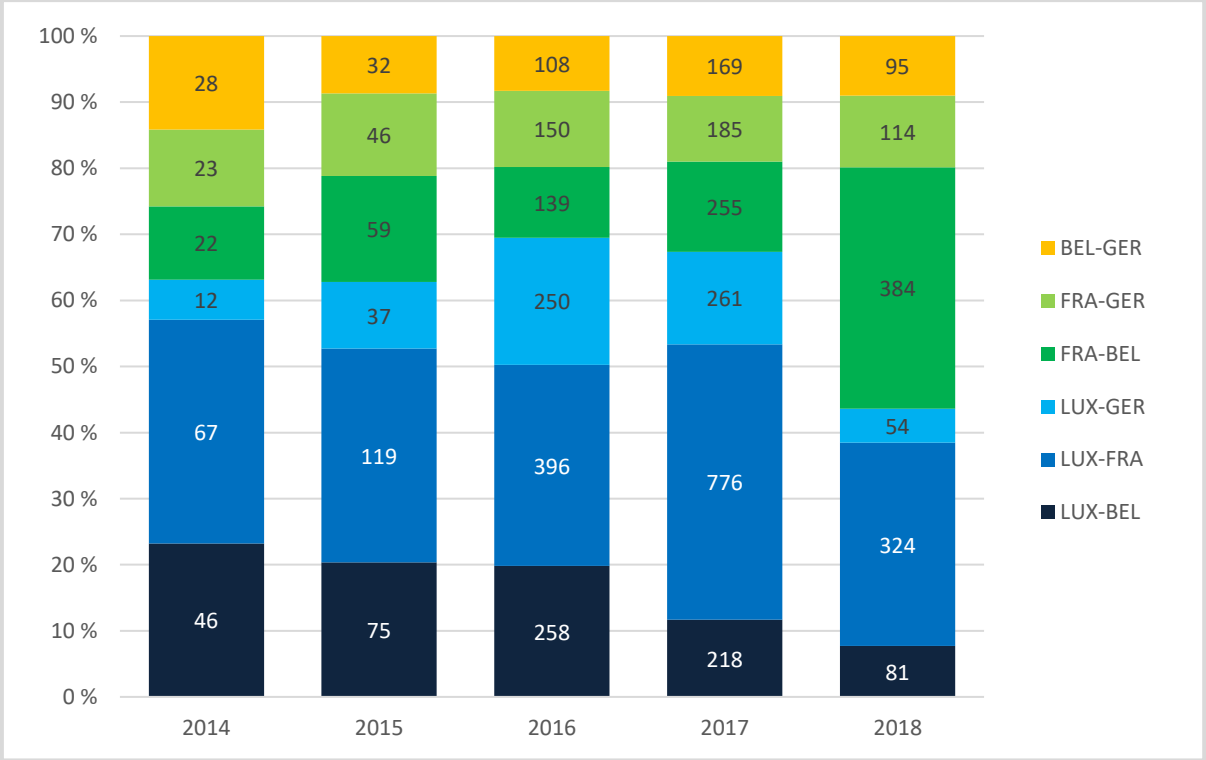


C-B both ways	2014	2015	2016	2017	2018
TOTAL (N)	628	964	1795	2424	1861

Figure 14. Cross-border movements (C-B) yearly in the Greater Region for the Greater Region users.

Firstly, cross-border movements are most dominant between France and Luxembourg both ways for individuals belonging to the Greater Region home region class (Figure 14). The relative share of cross-border movements between these two areas has also been progressively growing in recent years. In terms of other neighboring countries (Belgium and Germany), cross-border movements between Luxembourg and Germany are slightly more common than movements between Luxembourg and Belgium. However, the trend in both connections seems to be that their relative shares are dropping in relation to France-Luxembourg connection. Overall, cross-border movements including Luxembourg cover approximately 90 % of all trips made in 2014–2018.

POTENTIALS



C-B both ways	2014	2015	2016	2017	2018
TOTAL (N)	198	368	1301	1864	1052

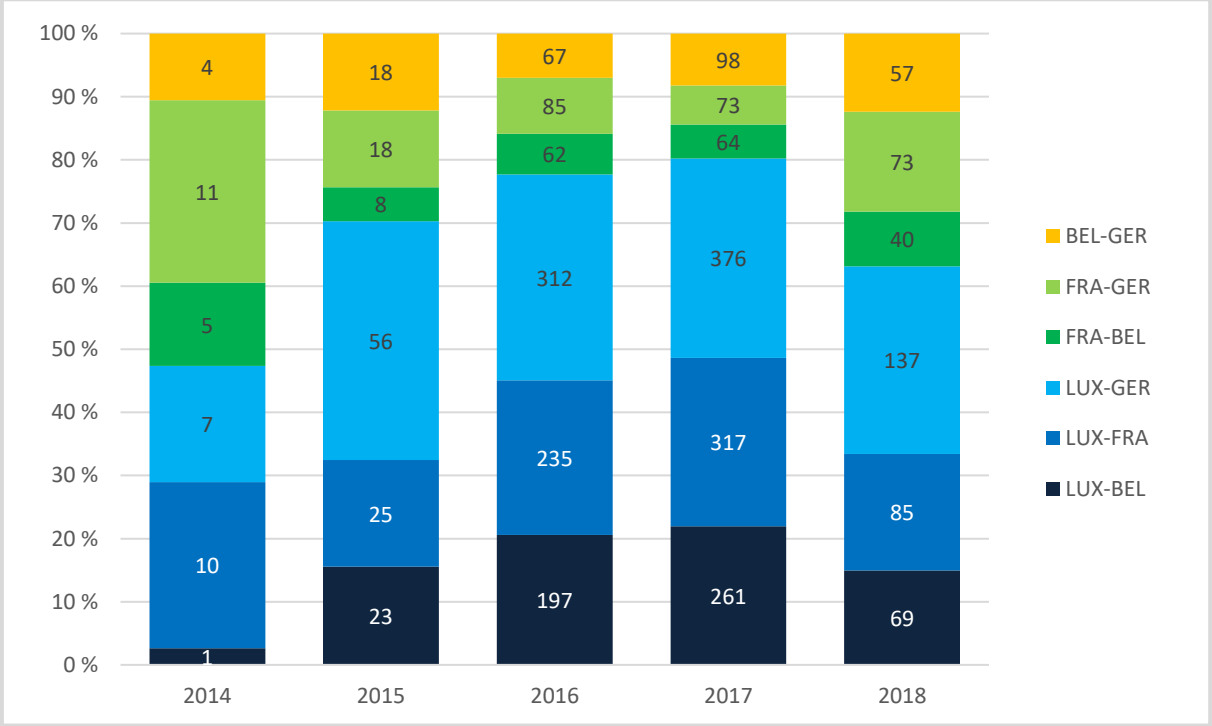
Figure 15. Cross-border movements (C-B) yearly in the Greater Region for Potentials.

Secondly, considering Potentials (Figure 15), one can clearly see the dynamics changing. Cross-border movements including Luxembourg are more infrequent; approximately 65 % of all trips in 2014–2017 and only slightly over 40 % in 2018. LUX-FRA movements, however, are still showing a strong interrelationship between the two countries. Movements between

Luxembourg's neighboring countries are also more common compared to the Greater Region class, especially in FRA-BEL.

In addition, considering the number of border crossings (i.e. total n-values), cross-border movements in the Greater Region are distinctly more infrequent for Potentials than users in the Greater Region group.

OTHERS



C-B both ways	2014	2015	2016	2017	2018
TOTAL (N)	38	148	958	1189	461

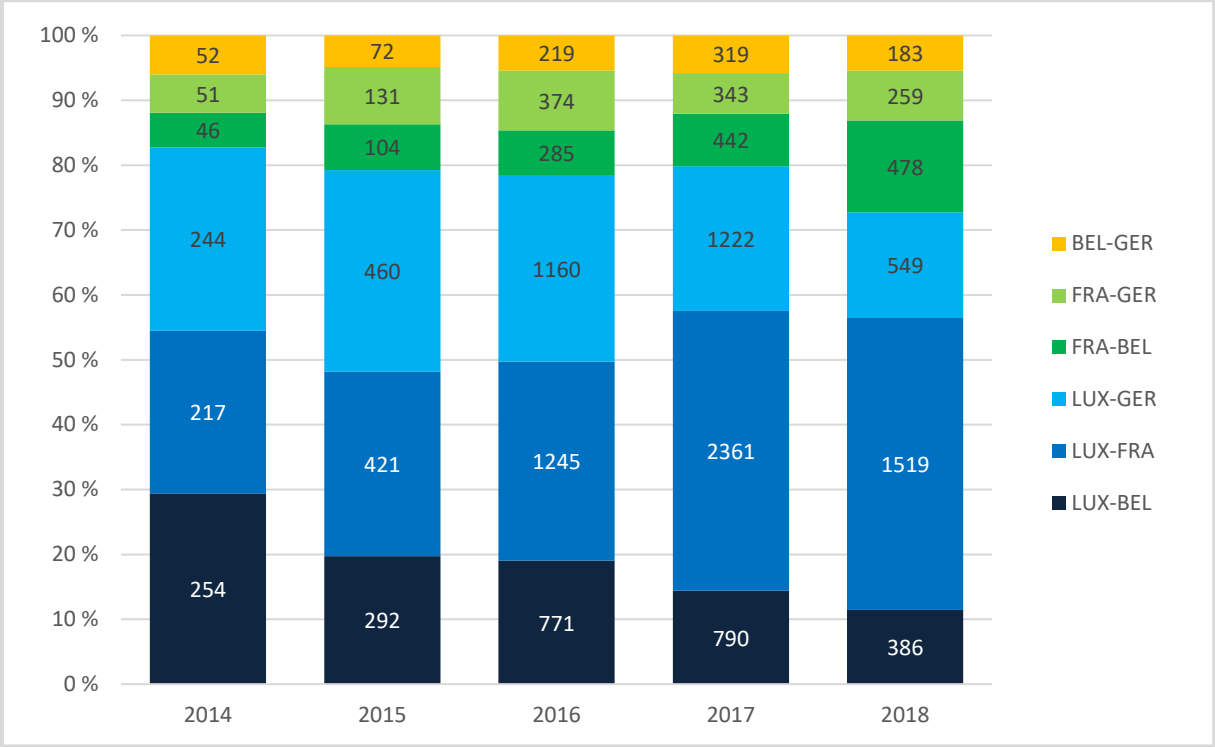
Figure 16. Cross-border movements (C-B) yearly in the Greater Region for Others.

Thirdly, cross-border movements in the Greater Region for Others are even more infrequent; only 2016 and 2017 have representative n-values (Figure 16). However, movements including Luxembourg and a neighboring country are relatively high, even up to 80 % in 2017. LUX-GER connection is distinctly the most dominant.

Finally, all trips combined, the relative share of cross-border movements including Luxembourg is approximately 80 % in 2014–2017 and circa 72 % in 2018 (Figure 17). The results are distinctly like the relative shares of the Greater Region users. The difference comes from BEL-GER, FRA-GER. and FRA-BEL trips. Their relative shares combined are close to

20 % whereas for the Greater Region users, it fluctuates between 8-15 %.

ALL



C-B both ways	2014	2015	2016	2017	2018
TOTAL (N)	864	1480	4054	5477	3374

Figure 17. All aggregated cross-border movements (C-B) yearly.

4.3 Defined and extracted cross-border mover types

4.3.1 Daily cross-border movers

A daily cross-border mover label was assigned to 172 users out of 702 (24.5 %) in the Greater Region based on the heuristic cross-border mover classification algorithm. Figure 18 shows the activity locations of daily cross-border movers inside the Greater Region with Luxembourg extracted.

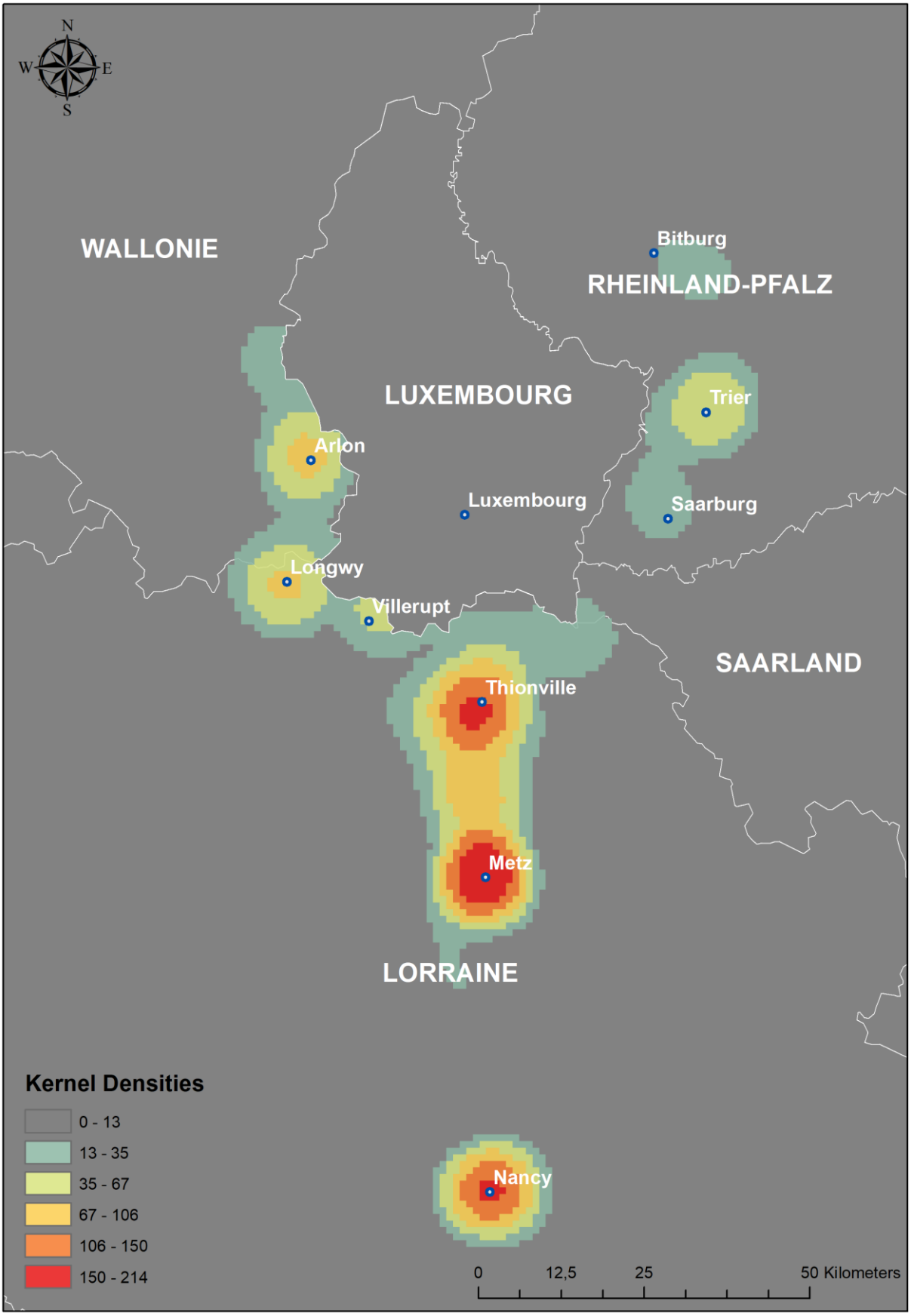


Figure 18. Daily cross-border mover activity location densities based on geotagged Twitter in 2010–2018. Kernel density cell size 1000 m^2, 10 000 m search radius. Luxembourg extracted. High density areas are situated in France.

The activity locations of the daily cross-border movers are distinctly located near the Luxembourg state boundary. The densest clusters are in France revealing Thionville, Metz, and Nancy as the hotspot areas. Other activity locations in France are Villerupt and Longwy, only a few kilometers away from the Luxembourg border in southwest.

In Belgium, the city of Arlon has the highest spatial density amongst daily cross-border movers, and in Germany, Trier, Saarburg, and Bitburg are the main activity clusters.

These activity areas are distinct start and endpoints for daily cross-border movements. Figure 19 represents all "Inside GRL" cross-border movements between dominance areas and Luxembourg both ways as line densities.

Firstly, Belgium-Luxembourg daily cross-border movements are spatially dispersed to several different areas. Yet, distinct high-density concentrations can be detected. Arlon-Luxembourg connection is distinguished as having the highest density. Other distinct links between the Grand Dutchy capital and Belgium include Martelange and Neufchàteau. Weaker connections to Jenneville, Libramont-Chevigny, Liége, and Namur also stand out.

In terms of other areas in Luxembourg; Clervaux, Differdange, and Mersch stand out having daily cross-border connections with Belgium. Clervaux has a distinct interconnection with both Oudler and Saint-Vith in the north, Differdange with Arlon, and Mersch with Jenneville.

Secondly, between France and Luxembourg, the overall spatial dispersion of the cross-border movements is much lower than between Belgium and Luxembourg. The main connections are distinctly Metz-Luxembourg and Thionville-Luxembourg. Other high-density concentrations include Differdange in Luxembourg. These daily cross-border movements cover Metz, Nancy, and Villerupt in France.

Finally, daily cross-border movements between Germany and Luxembourg are again more spatially dispersed. Movements including the Grand Dutchy capital are clearly connected to Trier and Bitburg - two of the three identified activity location clusters in Germany. Trier has high spatial connection also with Mersch, Diekirch, and Clervaux in Luxembourg, whereas Bitburg with Diekirch

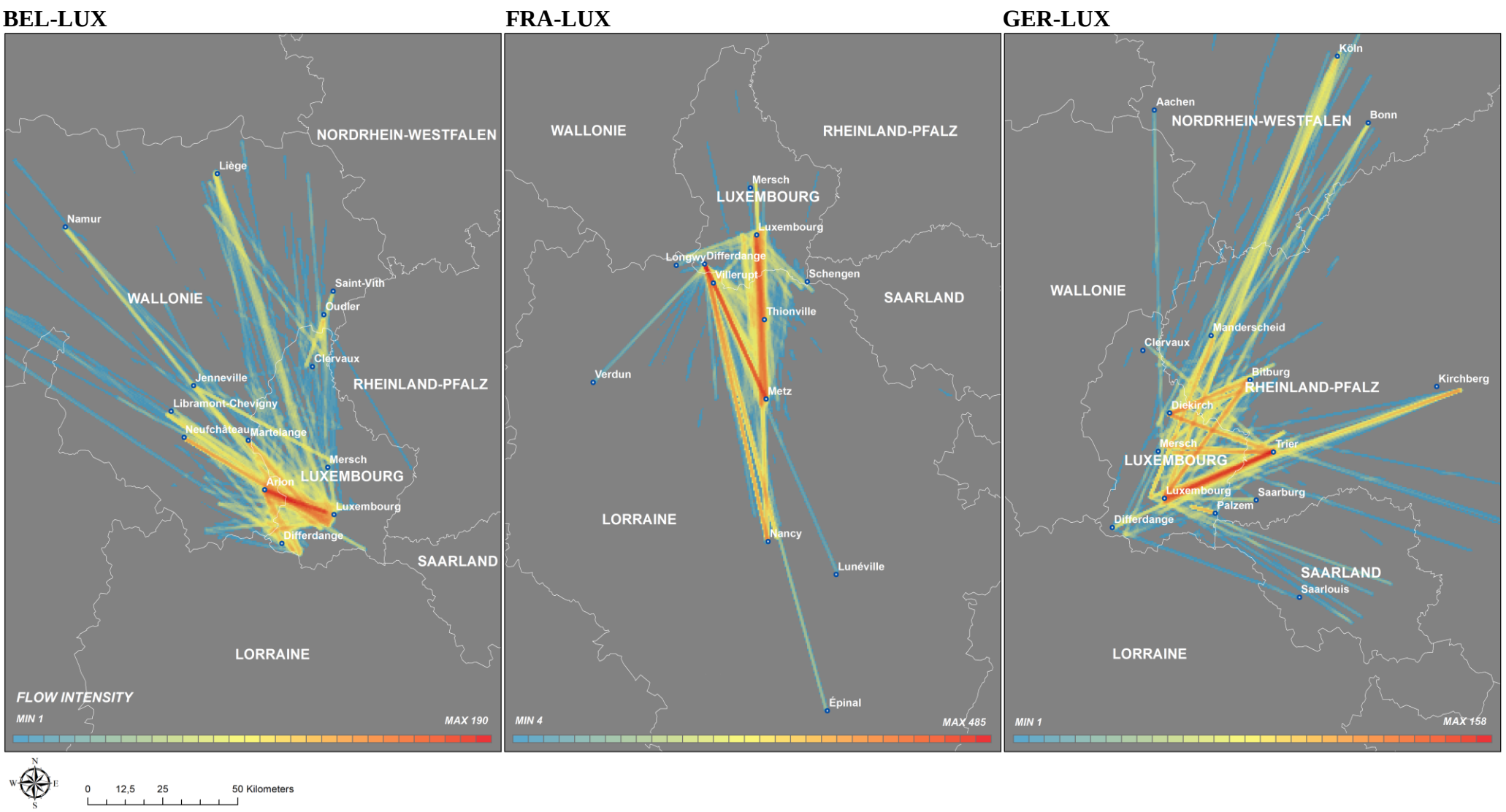


Figure 19. Inside GRL trips for identified daily cross-border movers in 2010–2018. Movements cover both ways between Luxembourg and surrounding dominance areas. Line density cell size 500 m^2, 750 m search radius. FRA-LUX has the highest maximum flow intensity, GER-LUX the lowest.

More remote areas are also identified in Nordrhein-Westfalen and Rheinland-Pfalz; Bonn and Köln are distinctly standing out as well as Kirchberg. They are all mainly connected to the city of Luxembourg.

Daily cross-border movements to Saarland reflect the overall Germany-Luxembourg connection – start and endpoints are relatively dispersed.

All in all, France-Luxembourg connections have the highest maximum flow intensity followed by Belgium-Luxembourg and Germany-Luxembourg.

4.3.2 Infrequent border crossers

While 172 users were identified as cross-border commuters, the rest of the Greater Region users (530 out of 702, 75.5 %) were classified as infrequent border crossers. Figure 20 shows the activity locations of these users inside the Greater Region with Luxembourg extracted.

In relation to the activity locations of the daily cross-border movers, infrequent border crossers have the same high-density area clusters in all dominance areas surrounding Luxembourg, but a lot more spatial dispersion can be detected, especially in Germany. Activity densities in Rheinland-Pfalz are distinctly diffused around Luxembourg's eastern boundary, and some activity centers can be detected in Saarland. In Belgium, densities expand slightly towards west, and the surroundings of Saint-Vith stand out as a local activity cluster. In France, the most apparent observation is that spatial densities around Thionville are distinctly lower and more dispersed than in the case of daily cross-border movers.

Considering the line densities for all "Inside GRL" trips between Luxembourg and surrounding dominance areas both ways for infrequent border crossers, the spatial dispersion becomes more evident than in the case of daily cross-border movers (Figure 21). Between Belgium and Luxembourg, Arlon-Luxembourg and Virton-Esch-sur-Alzette connections can be detected near the state boundary, but otherwise, cross-border movements seem to extend to more remote areas in northern parts of Wallonia. Between France and Luxembourg, the high-density cross-border movements appear to be relatively like the "Inside GRL" trips of the daily cross-border movers, but more infrequent movements are distinctly spatially dispersed to southern parts of Lorraine. Between Germany and Luxembourg, the situation resembles the Belgium-Luxembourg relation; one connection is clearly strong near the state boundary (Trier-Luxembourg) but mostly the cross-border movements are spatially extended to remote areas.

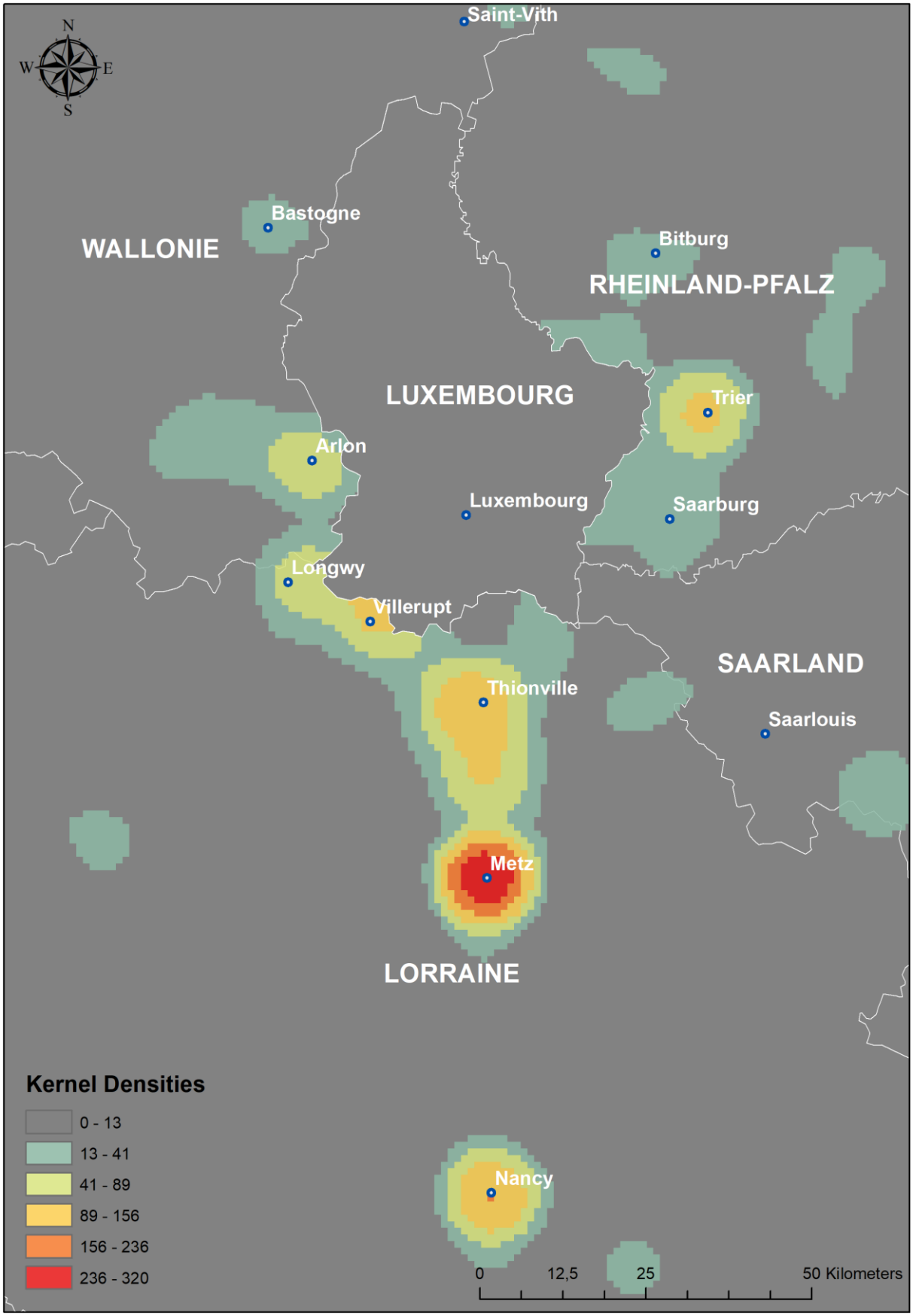


Figure 20. Infrequent border crosser activity location densities based on geotagged Twitter in 2010–2018. Kernel density cell size 1000 m^2, 10 000 m search radius. Luxembourg extracted. High densities are situated in France, but a lot more spatial dispersion can be detected in relation to daily cross-border mobilities.

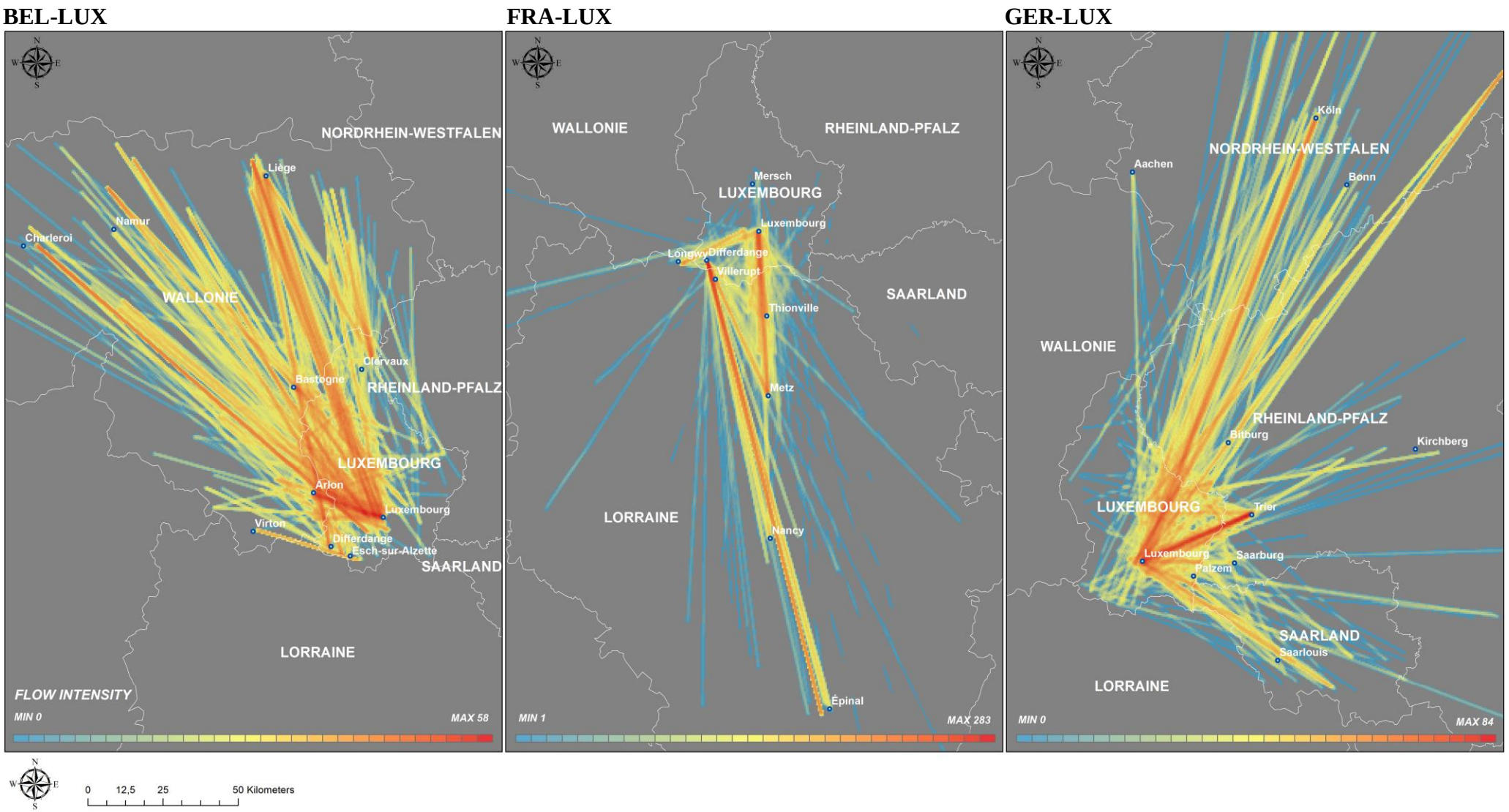


Figure 21. Inside GRL trips for infrequent border crossers in 2010–2018. Movements cover both ways between Luxembourg and surrounding dominance areas. Line density cell size 500 m^2, 750 m search radius. FRA-LUX has the highest maximum flow intensity, BEL-LUX the lowest.

4.3.3 Cross-border movement distances

Already comparing the "Inside GRL" movements of daily cross-border movers and infrequent border crossers (Figure 19 and Figure 21), it seems apparent that the two cross-border mover types differ from each other in terms of covered distances. Table 8 represents the average and median distances of "Inside GRL" movements for both mover types.

Table 8. Cross-border movement distance comparison between daily cross-border movers and infrequent border crossers. Year coverage 2010–2018.

	Distance AVG (km)			**Distance MDN (km)**		
CB-movement both ways	**BEL-LUX**	**FRA-LUX**	**GER-LUX**	**BEL-LUX**	**FRA-LUX**	**GER-LUX**
Daily cross-border mover	44	56	61	25	49	40
Infrequent border crosser	81	69	79	49	49	46
Difference	**37**	**13**	**18**	**24**	**0**	**6**

Considering the distance differences between the two mover types, an apparent observation is the difference between BEL-LUX daily cross-border movers and infrequent border crossers; both the average (37 km) and median (24 km) distance differences are relatively higher than the parallel FRA-LUX and GER-LUX values. GER-LUX connection is also divergent between the two cross-border mover types although the differential is not as strong as the difference in BEL-LUX; average distance difference is 18 km, and median difference 6 km.

The difference values between France and Luxembourg are slightly different from the two other dominance areas. The average difference is the lowest (13 km) and the median difference is neutral. This is also slightly apparent when considering the visual differences between cross-border movements; the high-density connections seem to be extremely close between daily cross-border movers and infrequent border crossers (Figure 19 and Figure 21).

In terms of individual daily cross-border connections, daily cross-border movers between Belgium and Luxembourg have relatively short trip distances; median value is only 25 km. However, distinct fluctuations are evident since the average trip distance is 44 km. Variations in daily cross-border movements can also be detected in Germany; median distance is 40 km,

average 61 km. In France, alterations seem to be slight; median value is 49 km and average 56 km.

Considering infrequent border crossers, the fluctuations are higher in all dominance areas. In Belgium, the average value is 81 km and median 49 km. The equivalent values in France are 69 km and 49 km, and in Germany 79 km and 46 km.

As an outcome, the trip distances covered by daily cross-border movers are evidently shorter than by infrequent border crossers. One can also state that the average values manifest bias due to individuals who travel long distances while median values represent central movement tendencies more accurately. Figure 23 represents these dynamics through cumulative distance graphs. It clearly shows that the median values are in a close correspondence to 50 % cumulative coverage. The only deviation to this is infrequent border crossers between Belgium and Luxembourg where the average value represents 50 % cumulative coverage in distances.

4.3.4 Temporal variation

The temporal variation inspections of cross-border movements were carried out on weekday level for both cross-border mover types. Considering daily cross-border movers, cross-border movements are predominantly occurring on business days whereas weekends' share of border crossings is way below average (Figure 22).

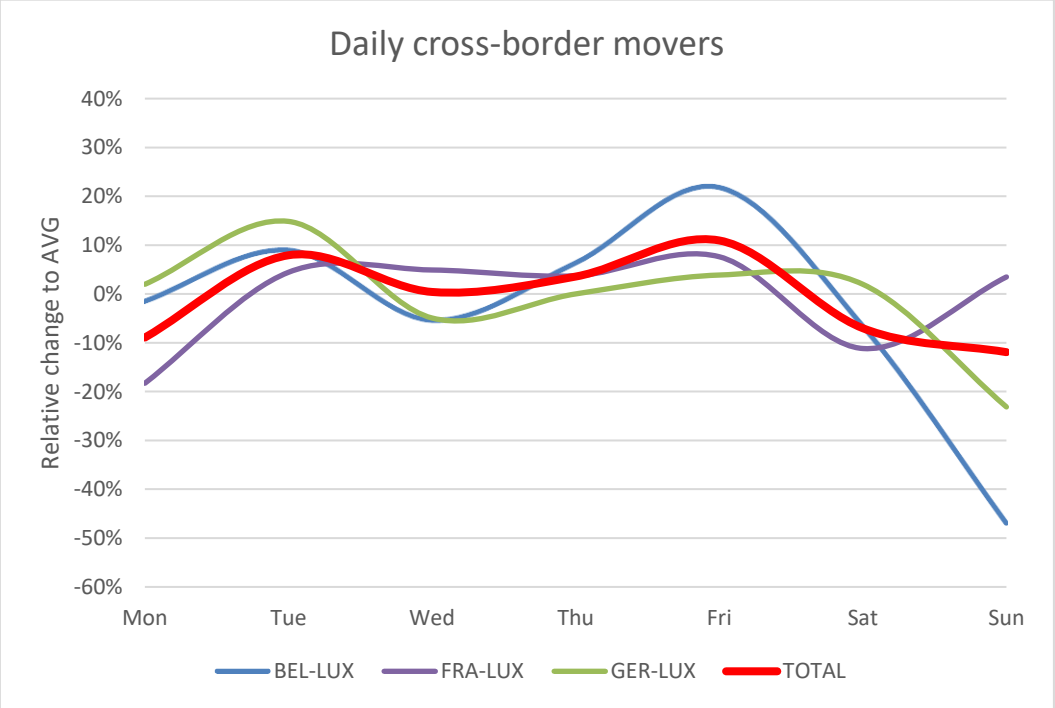
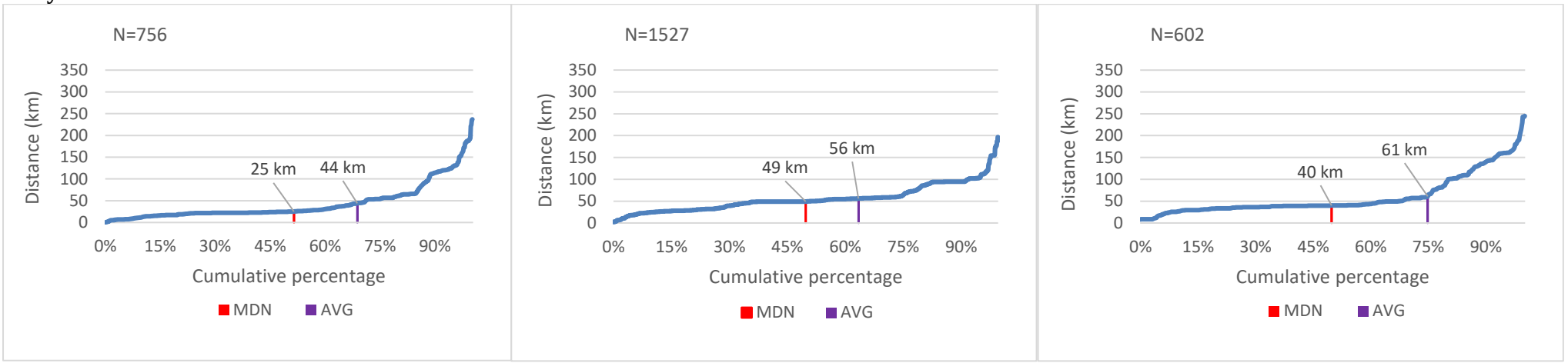


Figure 22. Weekday variation for daily cross-border movers. Inside GRL trips both ways in 2010–2018.

BEL-LUX **FRA-LUX** **GER-LUX**

Daily cross-border movers



Infrequent border crossers

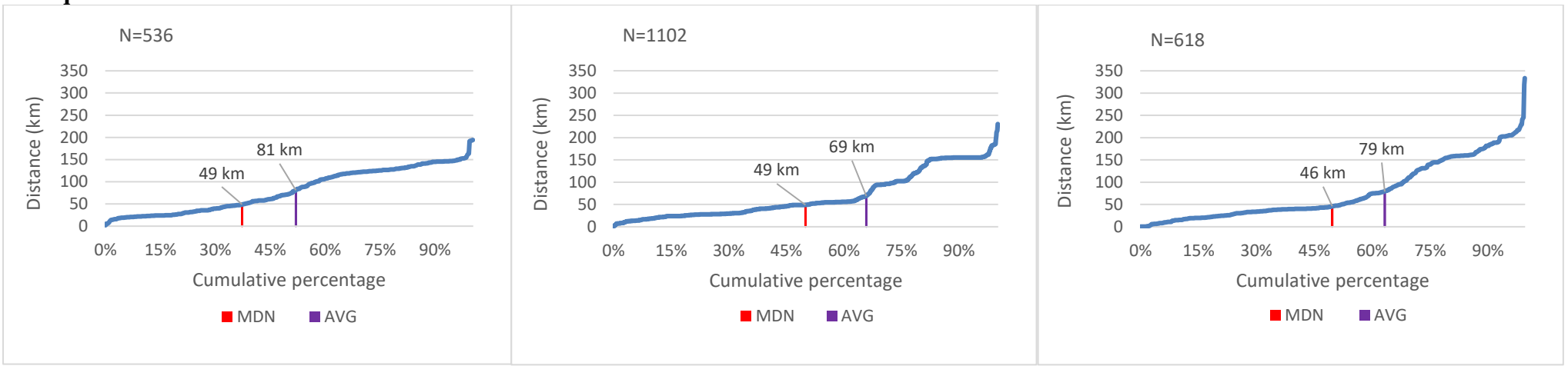


Figure 23. Inside GRL trip distances in 2010–2018. One can clearly see that individuals traveling long distances cause bias to average values.

The weekday variance for infrequent border crossers is de facto the opposite; the border crossings' share of occurrence on business days is distinctly below the average whereas on weekends, there is a clear peak on cross-border movements (Figure 24).

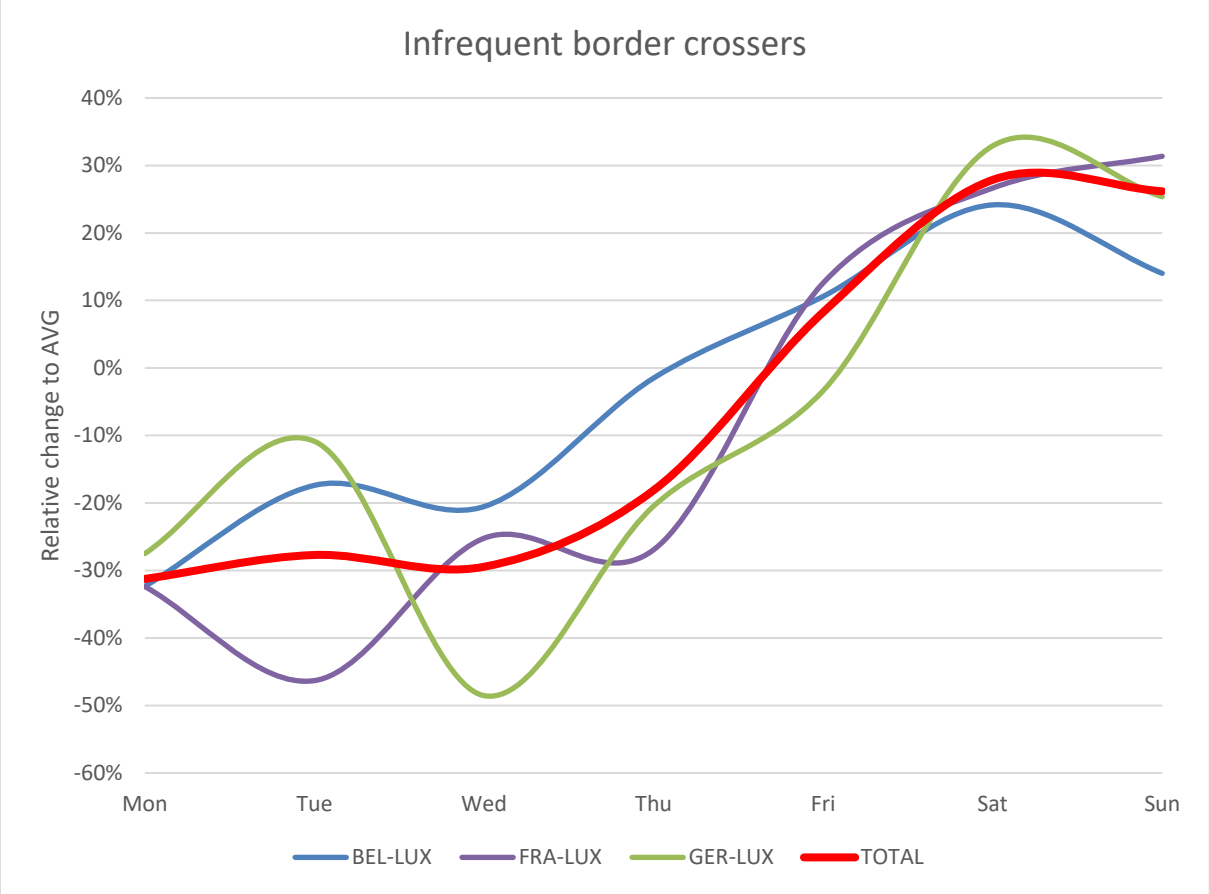


Figure 24. Weekday variation for infrequent border crossers. Inside GRL trips both ways in 2010–2018.

5. DISCUSSION

5.1 Data considerations

5.1.1 The coverage of geotagged Twitter depends on data acquisition processes

Previous studies (e.g. Carpentier, 2012; Gerber, 2012; Blanford *et al.*, 2015; Drevon *et al.*, 2016a) have shown that comprehensive data for cross-border mobility research has largely been missing; national statistics, registers, surveys, and census data have lacked information about the duration of activities. Partly due to this, investigations have focused mainly on aggregate levels, and person-based approach has been missing. There has been a distinct call for studying alternative data sources, including social media. In this study, geotagged Twitter data was utilized in the study area of the Greater Region of Luxembourg. Acquisition of the data was carried out on multiple levels.

Firstly, Digital Geography Lab from the University of Helsinki provided the initial dataset covering all tweets from 2016-2018 collected from the Twitter Streaming API on standard 1 % level for 124 994 users. Secondly, users who had posted at least once in Luxembourg with location information activated (i.e. geotagged post) were identified from the initial dataset. In retrospect, this could have covered the entire Greater Region of Luxembourg, not just the Grand Duchy, to achieve larger coverage. Thirdly, individual tweeting histories of identified users were collected from Twitter Search APIs' user timeline endpoint. This process resulted in a dataset in which geotagged tweet coverage was approximately 13.5 % for 3803 users (see Table 4).

In previous studies on location based social media, the representativeness of the data has received mixed opinions. Martí *et al.* (2019) conducted a literature overview on the matter stating that in some cases the sample sizes have been adequate to cover human activity on an aggregate level. However, in some cases, the samples' representativeness of the wider social media platform population has been questioned.

In previous studies on Twitter (Sloan *et al.*, 2013; Sloan and Morgan, 2015), it has been argued that users activating location information and publishing geotagged tweets are not representative of the wider Twitter population. As a reference, Sloan *et al.* (2013) reported a geotag coverage of 0.85 %, and Sloan and Morgan (2015) a relative share of 3.1 %. Both studies utilized Twitter Streaming API on standard 1 % level. In 2015, the dataset covered globally

more than 30 million tweeters during April.

The most distinctive difference between previous Twitter coverage research and this study in terms of data acquisition is the identification of users having location information activated. In this study, all users whose tweeting histories were collected from Twitter Search API had already activated their location information at some point (i.e. geotag identified). In previous studies (Sloan *et al.*, 2013; Sloan and Morgan, 2015), only Twitter Streaming API was utilized without filtering of location information activation.

The results show that this solution in data acquisition process increases geotagged tweet coverage considerably. Hence, in future studies, representativeness of geotagged Twitter should be considered through multi-level data acquisition processes. However, one must also consider the spatial coverage and n-values (i.e. user count). In this study, only a relatively small area was under main scrutiny (i.e. the Greater Region) while previous Twitter studies had a global spatial coverage. In addition, this study covered lower n-values; the initial dataset consisted of 124 994 users while previous studies had tens of millions.

Future studies should also consider content analysis in addition to spatio-temporal analysis approach. This could increase study sample and extend the spatial information of users from their digital footprints. For instance, geoparsing (i.e. extracting place names from text and returning location information) in relation to Twitter user profile could provide even better coverage. These types of options are already available e.g. for Python (Mordecai geoparsing library).

5.1.2 Twitter API highlights the most recent tweets causing yearly temporal bias

The Twitter dataset used in analysis had a temporal coverage of 2010-2018. However, as can be seen from the total n-values of aggregate cross-border movements yearly (Figure 14-Figure 17), the n-values are relatively lower in 2014 and 2015 than in 2016-2018. All years before 2014 were even lower covering only sporadic user tweeting activity (see Table 5). Hence, they were excluded from aggregate-level border crossing inspections.

This yearly bias was mainly caused by Twitter Search APIs' user timeline endpoint utilized in data acquisition. It returned 3200 most recent tweets for each individual user. If a user had less than 3200 tweets in the account, the API returned all tweets.

I carried out the user selection and filtering before using the Twitter search API, which meant

that this study was dealing with active Twitter users. Hence, it was natural that the user timeline endpoint largely returned posts from latest years.

For future studies, it is thus recommended to consider alternative Twitter API endpoints if a study aims to focus on tweeting activities in the past. For studies utilizing person-based approach based on active users, the user timeline endpoint gives valuable information; the tweets collected are not sporadic from different users, but systematically saved from each user at a time. In addition, this issue should be considered while evaluating the representativeness of the social media data in mobility research.

This study also suggests that a multi-year cross-border mobility analysis is possible to be carried out with user timeline endpoint including two- or three-years tweeting information. It is likely, however, that there would be effects on coverage if the Twitter Streaming API was used on either gardenhose or firehose level instead of standard (for differences in streaming levels, see 2.3.3 Social media data).

The coverage of the dataset is also affected by errors and anomalies related to data acquisition processes as well as filtering of unwanted information. In this study, the Twitter Search API did not return all tweeting histories of 4020 users. 217 users returned two types of errors; a 404-error meaning that the user was not found (i.e. removed from Twitter) and a 401-error meaning that the access was unauthorized (i.e. contents converted to non-public). In addition, 406 users did not return a single geotagged post although at least one was identified per user in the initial data. These contents might have been removed from Twitter or the user timeline endpoint could not reach these contents. For instance, if a user had posted a geotag on February 2016 but the user timeline endpoint returned only 3200 most recent posts between March 2016 and December 2018, the geotag coverage might have been zero for the user at question.

Furthermore, the user timeline endpoint returned 25 704 empty JSON strings (i.e. empty records having no attribute information about the collected tweets). Twitter server/API overloads or issues with packing the data as part of data acquisition process might have caused this. In future studies, it is thus recommended to use a database for data storage, an important remark also pointed out by Poorthuis and Zook (2017). The JSON format returned by the Twitter Search API is well suited for NoSQL databases, e.g. MongoDB. Using a database would precipitate the data collection process and decrease a risk for user-based errors.

If an object-relational or spatial database is used, a schema must be built separately. Then, it is

recommended to use an "adhoc" approach instead of an elaborate data collection framework used in this study. As a default, the Twitter API returns a wide variety of user profile information, which will complicate building of a database schema (Twitter, 2019).

Bot detection in this study was carried out using a 40 % threshold based on findings from previous studies (Hasnat and Hasan, 2018; Wojcik *et al.*, 2018). However, this issue has not been investigated much. As a result, the 40 % threshold should be considered critically in the future studies; some bots might not have been removed from the dataset, and conversely, some non-automated content might have been removed.

5.2 Methodological reflections

5.2.1 Twitter user profile can provide insights for home country detection

In recent studies, detecting home country accurately from non-continuous traces (i.e. social media and mobile phone data) has been reported to be extremely difficult (Bojic *et al.*, 2015; Hasnat and Hasan, 2018; Vanhoof *et al.*, 2018). The main issues have been that heuristic thresholds have been placed "blindly" without adequate reference (Vanhoof *et al.*, 2018) and that possibilities for validation have been limited (Hasnat and Hasan, 2018; Vanhoof *et al.*, 2018). Bojic *et al.* (2015) also point out that choosing the most suitable method should be considered individually for each dataset since datasets are unequally susceptible to varied methods. Hence, it is demanding to select the methods so that they serve each study setting in a most suitable manner.

A vast number of home detection algorithms presented for non-continuous traces in previous studies are leaning on either heuristics (e.g. Li *et al.*, 2012; McGee *et al.*, 2013; Hawelka *et al.*, 2014) or complex decision rules (e.g. machine learning solutions) (e.g. Ahas *et al.*, 2010; Frias-Martinez and Virseda, 2012). To the writer's knowledge, only Hasnat and Hasan (2018) are considering user profile information as the ground truth information for validation. However, this information was not utilized as part of home detection and assignment.

In this study, user-given home location was labeled as a home country if the information had been reported unambiguously and with enough spatial accuracy in user profile. Based on these criteria, I was able to detect and assign a home country for 33 % of all users. This relative share could have been extended if the user-given information had been applied to countries outside the Greater Region. However, this was irrelevant in the context of the selected study area and

study setting. It might have excluded users who had named a home country outside the Greater Region but who are, nonetheless, manifesting daily cross-border mobility in the area.

Later in the analysis, the user-given information was connected to heuristics when identifying dominance areas (i.e. daily life spaces of people) inside the Greater Region. Activity location densities (Figure 18 & Figure 20) seem to correspond well to previous studies (see 5.3.1) which also supports the relevance of utilizing user profile information as part of home detection.

Thus, this study argues that the Twitter user profile can provide insights for home country detection, not only for validation. If the user-given information for home country has not changed and the user-given location corresponds to the spatio-temporal activity of an individual, user profile can be argued to give insights to the user's activity locations also in a wider sense. However, this study did not expand the analysis to cover individual activity spaces as e.g. Järv *et al.* (2014) did, an issue proposed to be considered in the future studies.

Overall, future social media and home detection studies should give more emphasis to user content as well as advanced methods. Heuristics and complex decision rules connected to user-reported information can provide an asset to the selection of a suitable home detection method in each study setting and reveal new insights for human mobility.

5.2.2 Heuristic home country detection methods result in high accuracy on a regional level

Recently, Hasnat and Hasan (2018) compared different home detection methods in tourist identification via geotagged Twitter in terms of accuracy, precision, recall, and f-score. Adaptive boosting, a machine learning method, returned the best accuracy result of 82.4 %. One heuristic algorithm was also introduced (accuracy being 79.1 %) relying on time-based limitations for each night. All calculations were based on 108 560 geotagged tweets for 6519 users identified in Central Florida region. To the writer's knowledge, other similar inspections have not been conducted using social media data in the context of daily mobilities.

In this study, the developed "unique weeks" heuristic HDA returned an accuracy of 88.6 % in relation to the ground truth extracted from geotagged Twitter data, the "unique days" 87.1 %. The results in relation to previous ones are extremely promising and give valuable insights for home detection approaches in future research. However, it is recommended to consider also more advanced methods, e.g. complex decision rules. If the accuracy results were better in the context of tourist identification, the outcome in the context of daily movements could also be

even higher. Content analysis could also enhance both accuracy and precision.

These results should, however, be considered cautiously since many external factors are different in this study than in what Hasnat and Hasan (2018) conducted. Firstly, in tourist identification, one is trying to identify sporadic behavior from an individual perspective whereas daily cross-border mobilities are reoccurring movements in a specified catchment area. Hence, they are antonyms by nature and require different approaches in terms of home detection. Hasnat and Hasan (2018) labeled users who didn't satisfy the HDA's criteria as tourists, but in this study, the results from home detection were the foundation for daily cross-border mover extraction in the Greater Region area.

Secondly, in this study, home detection validation was carried out in the Greater Region of Luxembourg, an area located inland. Florida, on the other hand, is a peninsula mainly surrounded by the sea. In the Greater Region, an individual could easily have reoccurring activities in two or more administrative areas since every boundary is shared with another country inside the Schengen Area. In Florida, the nearby human activities can only be in Georgia or Alabama up north instead of Florida; there are less common administrative boundaries with other areas that in the Greater Region. Hence, the risk of spatial errors is higher in the Greater Region.

This aspect was considered in this study while assigning dominance areas for each user having most activities in the Greater Region. As can be seen from Table 6, majority of the users had the most tweeting activity inside Luxembourg, and Belgium seems to be underrepresented. It is possible that some users who are living e.g. in Belgium were classified to Luxembourg due to Twitter activities. Thus, all cross-border movements between Luxembourg and neighboring countries were analyzed covering both ways. This issue could be solved through content analysis of tweets to reveal actual home locations.

Finally, the fact that both validations were constricted to geographically distinct regional levels, it is still difficult to express whether conclusions can be made globally. This observation is in line with findings of Goodchild (2013) and Martí *et al.* (2019), who argue that individual case studies based on location-based social media include problems in terms of transferability of the results to other geographical areas. However, as already Hägerstrand (1970) has pointed out, "nothing truly general can be said about aggregate regularities until it has been made clear how far they remain invariant at micro level". Hence, in future home detection studies, it is

recommended to expand the geographical coverage to more global while still maintaining the person-based approach.

5.2.3 Counting border crossings accurately is challenging

Movement extraction in this study was based on ordering geotagged tweets per user in a chronological order. Two consecutive posts were interpreted to represent a trip if under 365 days was passed between the tweets. Both origin and destination countries were stored as attributes to identify cross-border movements between countries.

In previous studies, movement extraction from social media data has not been seen as a major predicament (Hawelka *et al.*, 2014; Blanford *et al.*, 2015). Basic assumption has been that consecutive posts represent a trip by nature, and only inconsistent speed of travel has been excluded (e.g. speed over 1000 km/h) (Hawelka *et al.*, 2014; Blanford *et al.*, 2015).

However, estimating the locations and activities between consecutive posts has been noted being challenging (Luo *et al.*, 2016). Luo *et al.* (2016) assumed that individuals stay at the same location in between two posts. Temporal filtering was implemented so that if a user's tweeting activity was less than one tweet per week, an individual was excluded. Mobility was studied inside the city of Chicago with six months temporal coverage (January 1st, 2013 – June 30th, 2013).

In this study, the same approach was difficult to implement. This study focused on total movement patterns from several years trying to identify daily cross-border movement patterns. In the context of cross-border mobility, if e.g. 150 days has been passed between two back-to-back tweets from different countries by an individual, it can be stated that a state boundary has been crossed. If temporal filtering had been implemented similarly than Luo *et al.* (2016) had, this would have excluded many state boundary crossings. Hence, I introduced a 365-day threshold.

Some challenges, however, do emerge if the temporal filtering is extended from one-week average to a maximum of 365 days. For instance, in the case of 150 days, it is impossible to say how many times the state boundary has been crossed although it has happened at least once. Inversely, it is also possible that an individual has crossed a state boundary although there has not been any social media activity. Hence, counting border crossings accurately jointly with movement extraction from geotagged Twitter is challenging. In addition, this raises a question

of the level of social media equivalence to human cross-border mobility.

To the writer's knowledge, these aspects have not been studied before, and it is thus recommended for future cross-border mobility studies based on social media to investigate these aspects more. In the case of Twitter, utilizing better streaming levels could also influence approximating border crossings more accurately since the tweeting coverage would be higher.

5.2.4 Heuristic cross-border mover algorithm provides a good starting point for future cross-border mobility studies

This study was, to the writer's knowledge, the first attempt to try to identify daily cross-border movements using geotagged Twitter data. Hence, direct parallels in terms of methods could not be found from previous studies to do comparison with.

Previous studies focusing on daily cross-border movements in the Greater Region of Luxembourg (Carpentier, 2012; Gerber, 2012; Drevon *et al.*, 2016a) have analyzed cross-border commuters based on questionnaires/surveys, registers and national statistics, not Big Data. Hence, geographical knowledge discovery for daily movers has not been needed. In terms of mobility studies utilizing geotagged Twitter (Hawelka *et al.*, 2014; Blanford *et al.*, 2015; Luo *et al.*, 2016), both the cross-border character and the extraction of daily mover types have not been under inspection. Thus, they do not provide a comprehensive starting point for daily cross-border mobility studies via social media.

Previous studies had introduced distance thresholds for cross-border commuter identification (Strüver, 2002; Gerber, 2012; Gerber, 2012 cit. Orfeuil, 2000). However, using distances as the basis in this study would have been inconsistent due to person-based approach; the mobility patterns of individuals fluctuate spatio-temporally (Järv *et al.*, 2014; Willberg, 2019). In other words, the trips of infrequent border crossers could also include short movements in terms of distances, and vice versa.

Thus, geographical knowledge discovery and sentiment analysis was conducted in this study since daily cross-border movements had to be separated from infrequent activities. The methods developed and introduced in sentiment analysis are recommended to be refined in future studies since this was only the first suggestion for the binary classification of cross-border movers. However, as the results show, the spatial dispersion of movements is distinctly lower amongst identified daily cross-border movers than with infrequent border crossers. Thus, it can be argued

that the heuristic cross-border mover algorithm developed and introduced in this study provides a valid starting point for future cross-border mobility studies utilizing social media.

5.3 The significance of the results in cross-border mobility research

5.3.1 Cross-border mobility patterns derived from social media data are in line with previous studies

Aggregate-level cross-border mobility patterns: Social media activity locations in the Greater Region are represented in Figure 12. The map clearly shows that the activities are mainly centered on both sides of France-Luxembourg state boundary. This outcome already supports the remark that there are cross-border movements happening in the area although the map only describes individual tweeting clusters instead of movements.

The France-Luxembourg dominance for individuals manifesting most activities in the Greater Region in terms of aggregate cross-border movements yearly is being revealed in Figure 14. It clearly shows that France-Luxembourg connections both ways have been growing while Germany-Luxembourg and Belgium-Luxembourg connections have lost their relative share due to the dominance of the France-Luxembourg cross-border movements. Germany-Luxembourg and Belgium-Luxembourg connections are close to one another in terms of numbers, Germany being slightly more dominant.

These outcomes are being supported by the official statistics (STATEC, 2016); Figure 1 distinctly shows that daily cross-border activities between Luxembourg and France have been the highest in the area. In second place comes Belgium, and Germany is third, although the two countries are almost even. Considering that this study included Nordrhein-Westfalen in Germany as part of the Greater Region, it is expected that Germany has more cross-border connections than Belgium to Luxembourg. If the relative share of Nordrhein-Westfalen was to be removed, the cross-border movements would correspond to official statistics even more accurately.

The results are also equivalent to previous empirical studies (Carpentier, 2012; Gerber, 2012). Gerber (2012) reported that both in 1995 and 2007 Luxembourg received most of the cross-border commuters from France, Belgium, and Germany, in that order. In terms of cross-border movement progression, Gerber (2012) reported a progression (%) of 130.6 between France-Luxembourg, 108.6 between Belgium-Luxembourg, and 193.9 between Germany-Luxembourg from 1995 to 2007. It seems that the progression of France-Luxembourg has continued to grow

while Belgium-Luxembourg and Germany-Luxembourg have dropped slightly. The relative decrease of Belgium-Luxembourg cross-border connections with respect to Germany and France has been reported by Carpentier (2012).

In terms of Potentials (Figure 15) and Others (Figure 16), two main conclusions can be made. Firstly, the results give a good reference to a valid dominance area detection inside the Greater Region; living farther away from Luxembourg in neighboring countries reduces the number of cross-border trips made to the Grand Dutchy. Secondly, although this study didn't focus on cross-border mobilities between two Luxembourg's neighboring countries (e.g. France-Belgium), some patterns can be detected that respond to previous studies. Gerber (2012) reported a France-Belgium progression of 169.5 % from 1997 to 2005. Considering especially Figure 15, the France-Belgium interconnection seems to have grown rapidly between 2017 and 2018.

Furthermore, the results from aggregate-level pattern inspections also give valuable insights to extracting border crossings from social media data. Although it is demanding to accurately approximate how many times an individual has crossed a state boundary between two consecutive posts, the results seem to indicate that movement extraction from geotagged Twitter based on user timeline endpoint provides applicable generalization for multi-year country-level inspections.

Daily cross-border mobilities: Considering activity location densities identified in this study (Figure 18) in relation to previous studies (Figure 3), one can clearly see that the highest density clusters correspond to one another. The main difference is that in this study Metz has higher density than Thionville. This, however, could be explained by population densities (see 5.3.3) which in Metz is much higher than in Thionville.

The results also correspond to previous studies in terms of cross-border movements. Comparing results in this study (Figure 19) and results from Drevon *et al.* (2016a) (Figure 4), one can clearly see that the highest line densities identified in this study correspond to activity spaces modeled previously. In addition, identified daily cross-border mobilities provide new information about the spatial extent of the movements; new inter-regional connections are revealed e.g. in Martelange-Luxembourg (Belgium-Luxembourg connections), Metz-Differdange and Nancy-Differdange (France-Luxembourg connections), as well as Trier-Mersch (Germany-Luxembourg connections). However, one must remember that these results

are compared to cross-border commuting. In this study, the nature of daily cross-border mobilities wasn't considered; some cross-border movements might not represent cross-border commuting but something different.

In terms of distances, Gerber (2012) stated that cross-border commuting trip distances can be relatively low on average, even down to 40 km or lower. Schmitz *et al.* (2012) reported that for France-Luxembourg cross-border commuters the average trip distance was 40 km in 2010. Equivalent value for Germany-Luxembourg was 46 km and for Belgium-Luxembourg 49 km.

Considering Table 8, some variation to these numbers can be detected. However, if median distances are considered, the results are in close correspondence to previous studies. In Belgium-Luxembourg, the average trip distance (44 km) is extremely close to previous studies but for France-Luxembourg and Germany-Luxembourg movements the average values are distinctly higher. However, Nordrhein-Westfalen was included in Germany, which clearly elevates the average distance. Considering the median value (40 km), the effect of Nordrhein-Westfalen decreases, and the distance corresponds to previous reports. France-Luxembourg connections higher average values are being explained by movements including Nancy, Lunéville, and Èpinal (Figure 19); cities not covered in previous studies in terms of cross-border commuting (Figure 3 & Figure 4).

It is, however, important to point out that it is possible that the heuristic cross-border mover algorithm developed in this study emphasizes infrequent activities in some areas. Hence, comparing distances to previous studies should be considered with a slight critical stance. Nonetheless, the results from temporal variation inspections give interesting insights about the issue (see 5.3.2).

Future studies: In this study, all movement patterns were analyzed both ways in relation to Luxembourg. In future studies, cross-border movement patterns should also be investigated centrifugally and centripetally to better understand the orientation of the movements in relation to Luxembourg. This remark has also been pointed out previously by Carpentier (2012): "if these centrifugal movements to Luxembourg are now relatively well known, we still understand very little about the centripetal movements."

In addition, cross-border connections between France and Belgium, France and Germany, as well as Germany and Belgium should be given more attention. Figure 1 and Figure 15 clearly show that there are interconnections between these areas. Also, the Potentials group still has

high connections to Luxembourg. Hence, it is recommended to consider expanding the spatial extent of daily cross-border mobility inspections to cover all of Luxembourg's neighboring country areas, not just the Greater Region.

Finally, future studies should also consider the social aspect of daily cross-border mobilities; some movements might represent cross-border commuting, but how universal this is in a wider perspective?

5.3.2 Temporal patterns and distance variations reveal different types of cross-border movements

This study concentrated only slightly on temporal aspects of cross-border movements in the Greater Region – the focus was on spatial movement patterns from several years. However, temporal variation on weekday-level was carried out. The results are extremely promising and give validation to the heuristic cross-border mover algorithm developed and used. Considering cross-border commuting as a phenomenon, people are usually working on weekdays. The results clearly show that identified daily cross-border movers are crossing a state boundary more frequently on weekdays (positive variance) than on weekends (negative variance) (Figure 22). Hence, it is likely that many of them represent cross-border commuting. When it comes to infrequent border crossers, the outcome is the opposite; weekdays show less frequent border crossings, but weekends are on the positive side of variance (Figure 24).

In future studies, these findings could be used in e.g. identification of cross-border commuters. Weekdays could be given a higher weighting than weekends in heuristic algorithms, which could result in more precise results. In addition, temporal variation could be detected on different levels (e.g. months) to better understand the spatio-temporal fluctuations of daily cross-border movements. In terms of movement distances (Table 8), there are distinct differences in covered distances between the two mover types. This gives justification for separating daily cross-border mobilities from other cross-border movements and manifests the validness of Tobler's first law of geography; "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970).

5.3.3 Weighting data with population density and Twitter use activity should be considered

Although the results presented in this study respond to previous cross-border mobility studies

utilizing surveys and national statistics, some critical stance is recommended; it is justifiable to say that both population densities and tweeting activities in the Greater Region countries influence geotagged Twitter's coverage of the wider population. For instance, on a country level, high population densities might partly cause the dominance of France in terms of spatial activity. Again, on a city level, it is known that Metz has a much higher population density than Thionville, which could explain why in this study Thionville does not appear to have as much daily cross-border movements as Metz has.

High population density indicates higher potential for social media activities since more people are in a specific area. The results should thus also be considered in relation to tweeting activities. This information, however, is challenging to find in a comprehensive form. Table 9 represents Twitter penetration rates in 2019 for the Greater Region countries.

Table 9. Twitter penetration rates in the Greater Region countries in 2019 according to data report by Kepios (2019).

	Twitter penetration rates			
Year	Luxembourg	Belgium	France	Germany
2019	18.4 %	8.9 %	8.5 %	4.7 %

Luxembourg has distinctly highest Twitter penetration rate in relation to other countries. Belgium and France are close to one another with over 8 % but in Germany the tweeting activity seems to be relatively low. However, one must consider the special nature of the Greater Region when valuating these outcomes. A lot of cross-border commuting is occurring in the area, which means that tweeting activities are intermingled between different countries. In addition, Twitter use activity fluctuates temporally. Hence, an accurate presentation on how much effect these numbers have on the results and what the equivalence of Twitter to cross-border mobility is, is difficult to evaluate.

In future studies, it is recommended to consider population densities and social media use activities as weighting values to normalize the bias. In addition, more emphasis should be given to the nature of individual's Twitter usage behavior; in what kind of situations is Twitter used in relation to other social media platforms? This could give more insights to evaluating the equivalence of Twitter to cross-border mobility.

6. CONCLUSIONS

This study has shown that social media can be implemented in cross-border mobility research, and social media Big Data can provide a relatively good proxy for cross-border mobility of people on a regional level. However, to draw universal conclusions about the global cross-border mobility characteristics, further studies are needed. Other geographical extents and study settings might reveal something completely new and different. Nevertheless, the methods and tools developed in this study provide a valid starting point for future research on cross-border mobility.

Geotagged Twitter linked to user profile information revealed spatio-temporal cross-border mobility patterns in the Greater Region of Luxembourg. Utilizing heuristic programmatic approach, daily cross-border mobilities were able to be defined and extracted from social media Big Data with a close correspondence to previous studies. This study also produced additional information about the spatial extent of the movements; new inter-regional connections were revealed. Thus, this study offers a good starting point for future daily cross-border mobility studies utilizing social media. In addition to mobility research, this study offers valuable insights to future research on home detection utilizing semi-continuous data traces, e.g. social media or mobile phone data.

There are still issues related to extracting cross-border movements from geotagged Twitter; estimating the number of border crossings in between consecutive posts is challenging. However, the results seem to indicate that movement extraction from geotagged Twitter based on user timeline endpoint provides applicable generalization for multi-year regional-level inspections.

To the writer's knowledge, this study was one of the first attempts in trying to identify daily cross-border mobilities using social media Big Data, hence heuristic programmatic approach. All scripts used are openly available on Digital Geography Lab's GitHub-pages (https://github.com/DigitalGeographyLab/cross-border-mobility-twitter). Critical stance and refinement of these newly developed methods is highly recommended.

ACKNOWLEDGEMENTS

First, I want to thank my supervisors Tuuli Toivonen and Olle Järv. All the support, comments, and overall guidance helped me enormously throughout the process. I want to point out Olle's massive role in helping me dive into cross-border mobility and making sense of the whole entirety. I want to thank Tuuli for helping me select a topic that was both challenging and rewarding.

Big thank you to all the great people in Digital Geography Lab for support and comments. Some of the methodological solutions were highly influenced by your insights. Especially I wish to thank Henrikki Tenkanen in assisting me getting the data acquisition process started.

Thank you, Spatial Needs, my musical comrades. Our training sessions and gigs made it easy for me to keep the balance between work and spare time.

An enormous thank you to my mom and my brother. The latest year has not been the easiest for us, but your support gave me strength to carry on with the work and ultimately get things done.

Finally, I want to thank my late father, a man who taught me everything about work ethics and how to be a responsible man. More than anything, I wish I could have shared this moment with you. I find comfort in knowing you would have been proud of me today.

LITERATURE

Ahas, R., Silm, S., Järv, O., Saluveer, E. and Tiru, M. (2010). 'Using mobile positioning data to model locations meaningful to users of mobile phones', *Journal of Urban Technology*, 17(1), pp. 3–27. doi: 10.1080/10630731003597306.

Blanford, J. I., Huang, Z., Savelyev, A. and MacEachren, A. M. (2015). 'Geo-located tweets. Enhancing mobility maps and capturing cross-border movement', *PLoS ONE*, 10(6), pp. 1–16. doi: 10.1371/journal.pone.0129202.

Bojic, I., Massaro, E. and Belyi, A. (2015). 'Choosing the right home location definition method for the given dataset', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9471, pp. 194–208. doi: 10.1007/978-3-319-27433-1_14.

Brunet-Jailly, Emmanuel (2011). 'Special Section: Borders, Borderlands and Theory: An Introduction', *Geopolitics*, 16(1), pp. 1–6. doi: 10.1080/14650045.2010.493765.

Calabrese, F., Ferrari, L. and Blondel, V. D. (2014). 'Urban Sensing Using Mobile Phone Network Data: A Survey of Research', *ACM Computing Surveys*, 47(2), pp. 1–20. doi: 10.1145/2655691.

Carpentier, Samuel (2012). 'Cross-border local mobility between Luxembourg and the walloon region: An overview', *European Journal of Transport and Infrastructure Research*, 12(2), pp. 198–210.

Castells, Manuel. (2000). *The rise of the network society*. Oxford, Blackwell Publishers.

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J-C., Huens, E., Van Dooren, P., Smoreda, Z. and Blondel, V.D. (2013). 'Exploring the mobility of mobile phone users', *Physica A: Statistical Mechanics and its Applications*, 392(6), pp. 1459–1473. doi: 10.1016/j.physa.2012.11.040.

Database of Global Administrative Areas (2019). *GADM maps and data*. Available at: https://gadm.org/download_world.html (Accessed: 6 September 2019).

Drevon, G., Gerber, P., Klein, O. and Enaux, C. (2016a). 'Measuring Functional Integration by Identifying the Trip Chains and the Profiles of Cross-Border Workers: Empirical Evidences from Luxembourg', *Journal of Borderlands Studies*. Taylor & Francis, 33(4), pp. 549–568. doi: 10.1080/08865655.2016.1257362.

Drevon, G., Gwiazdzinski, L. and Gerber, P. (2016b). 'Bricolages et arrangements des ménages dans les parcours de mobilité quotidienne', *Financé par le Fonds National de la Recherche Luxembourg (FNR)*.

Esri (2017). *Distance on a sphere: The Haversine Formula | GeoNet*. Available at: https://community.esri.com/groups/coordinate-reference-systems/blog/2017/10/05/haversine-formula (Accessed: 9 April 2019).

European Commission (2019). *Champagne-Ardenne - Internal Market, Industry, Entrepreneurship And Smes - European Commission*. Available at: https://ec.europa.eu/growth/tools-databases/regional-innovation-monitor/base-profile/champagne-ardenne (Accessed: 5 April 2019).

Frias-Martinez, V., Virseda, J., Rubio, A. and Frias-Martinez, E. (2010). 'Towards large scale technology impact analyses: automatic residential location from mobile phone-call data', pp. 1–10. doi: 10.1145/2369220.2369230.

Frias-Martinez, V. and Virseda, J. (2012). 'On the relationship between socio-economic factors and cell phone usage', p. 76. doi: 10.1145/2160673.2160684.

Gerber, Philippe (2012). 'Advancement in conceptualizing cross-border daily mobility: The Benelux context in the European union', *European Journal of Transport and Infrastructure Research*, 12(2), pp. 178–197.

Goodchild, Michael F. (2013). 'The quality of big (geo)data', *Dialogues in Human Geography*, 3(3), pp. 280–284. doi: 10.1177/2043820613513392.

Hasnat, M. M. and Hasan, S. (2018). 'Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data', *Transportation Research Part C: Emerging Technologies*. Elsevier, 96(September), pp. 38–54. doi: 10.1016/j.trc.2018.09.006.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. and Ratti, C. (2014). 'Geo-located Twitter as proxy for global mobility patterns', *Cartography and Geographic Information Science*. Taylor & Francis, 41(3), pp. 260–271. doi: 10.1080/15230406.2014.890072.

Herzog, L. A. and Sohn, C. (2016). 'The co-mingling of bordering dynamics in the San Diego–Tijuana cross-border metropolis', *Territory, Politics, Governance*, pp. 1–24. doi: 10.1080/21622671.2017.1323003.

Housley, W., Procter, R., Edwards, A., Burnap, P., Williams, M., Sloan, L., Rana, O., Morgan, J., Voss, A. and Greenhill, A. (2014). Big and broad social data and the sociological imagination: A collaborative response. *Big Data & Society*. https://doi.org/10.1177/2053951714545135.

Van Houtum, Henk (2005). 'The Geopolitics of Borders and Boundaries', *Geopolitics*, 10(4), pp. 672–679. doi: 10.1080/14650040500318522.

Hu, T., Luo, J., Kautz, H. and Sadilek, A. (2016). 'Home Location Inference from Sparse and Noisy Data: Models and Applications', *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*, 17(5), pp. 1382–1387. doi: 10.1109/ICDMW.2015.149.

Huang, A., Gallegos, L. and Lerman, K. (2017). 'Travel analytics: Understanding how destination choice and business clusters are connected based on social media data', *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 77, pp. 245–256. doi: 10.1016/j.trc.2016.12.019.

Huber, P. and Nowotny, K. (2011). 'Moving across Borders: Who is Willing to Migrate or to Commute?', *Regional Studies*, 47(9), pp. 1462–1481. doi: 10.1080/00343404.2011.624509.

Hägerstrand, Torsten (1970). 'What About People in Regional Science?', *Papers of the Regional Science Association*, 24(1), pp. 7–24. doi: 10.1111/j.1435-5597.1970.tb01464.x.

Hägerstrand, Torsten (1992). 'Mobility and transportation - are economics and technology the only limits?', *Facta and Futura*, 2, pp. 35–38.

Indiana University (2019). *Botometer® by OSoMe*. Available at: https://botometer.iuni.iu.edu/#!/ (Accessed: 6 April 2019).

Järv, O., Ahas, R. and Witlox, F. (2014). 'Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records', *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 38, pp. 122–135. doi: 10.1016/j.trc.2013.11.003.

Kaufmann, Vincent (2000). *Mobilité quotidienne et dynamiques urbaines: la question du report modal*. Presses polytechniques et universitaires romandes. Science, technique, société. Available at: https://books.google.fi/books?id=VRe6AAAACAAJ.

Kaufmann, Vincent (2005). 'Mobilités et réversibilités : Vers des sociétés plus fluides ?', *Cahiers internationaux de sociologie*, 118. doi: 10.3917/cis.118.0119.

Kaufmann, Vincent (2011). *Rethinking the city : urban dynamics and motility*. 1st ed. Routledge Milton Park, England.

Kellerman, Aharon (2012a). 'Chapter 1: Introduction', in *Daily Spatial Mobilities: Physical and Virtual*, ASHGATE, pp. 1–18.

Kellerman, Aharon (2012b). 'Chapter 2: Needs and Triggers for Daily Spatial Mobilities', in *Daily Spatial Mobilities: Physical and Virtual*, ASHGATE, pp. 21–36.

Kepios (2019). *Complete Report Library*. Available at:

https://datareportal.com/library (Accessed: 29 May 2019).

Kitchin, R. and McArdle, G. (2016). 'What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets', *Big Data & Society*, 3(1), p. 205395171663113. doi: 10.1177/2053951716631130.

Kung, K. S., Greco, K., Sobolevsky, S. and Ratti, C. (2014). 'Exploring universal patterns in human home-work commuting from mobile phone data', *PLoS ONE*, 9(6). doi: 10.1371/journal.pone.0096180.

Li, R., Wang, S., Deng, H., Wang, R. and Chang, K. C-C. (2012). 'Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations', *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining*, pp. 1023–1031. doi: 10.1145/2339530.2339692.

Luo, F., Cao, G., Mulligan, K. and Li, X. (2016). 'Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago', *Applied Geography*. Elsevier Ltd, 70, pp. 11–25. doi: 10.1016/j.apgeog.2016.03.001.

Lupton, Deborah (2015). *The thirteen Ps of big data | This Sociological Life, The Sociological Life*. Available at: https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/ (Accessed: 20 April 2019).

Mahmud, J., Nichols, J. and Drews, C. (2014). 'Home Location Identification of Twitter Users', *ACM Transactions on Intelligent Systems and Technology*, 5(3), pp. 1–21. doi: 10.1145/2528548.

Manca, M., Boratto, L., Roman, V. R., Gallissà, O. M. and Kaltenbrunner, A. (2017). 'Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study', *Online Social Networks and Media*. Elsevier B.V., 1, pp. 56–69. doi: 10.1016/j.osnem.2017.04.002.

Martí, P., Serrano-Estrada, L. and Nolasco-Cirugeda, A. (2019). 'Social Media data: Challenges, opportunities and limitations in urban studies', *Computers, Environment and Urban Systems*. Elsevier, 74, pp. 161–174. doi: 10.1016/j.compenvurbsys.2018.11.001.

McGee, J., Caverlee, J. and Cheng, Z. (2013). 'Location prediction in social media based on tie strength', pp. 459–468. doi: 10.1145/2505515.2505544.

Melakessou, F., Derrmann, T. and Engel, T. (2015). 'Asymmetry Analysis of Inbound/Outbound Car Traffic Load distribution in Luxembourg', pp. 5–12. doi: 10.1145/2810362.2810374.

Miller, Harvey J. (2017). 'Time Geography and Space-Time Prism', *International*

Encyclopedia of Geography, pp. 1–19. doi: 10.1002/9781118786352.wbieg0431.

Paasi, Anssi (1999). 'Boundaries as social practice and discourse: The Finnish-Russian border', *Regional Studies*, 33(7), pp. 669–680. doi: 10.1080/00343409950078701.

Paasi, A. and Prokkola, E.-K. (2008). 'Territorial Dynamics, Cross-border Work and Everyday Life in the Finnish–Swedish Border Area', *Space and Polity*. Routledge, 12(1), pp. 13–29. doi: 10.1080/13562570801969366.

Phithakkitnukoon, S., Smoreda, Z. and Olivier, P. (2012). 'Socio-geography of human mobility: A study using longitudinal mobile phone data', *PLoS ONE*, 7(6), pp. 1–9. doi: 10.1371/journal.pone.0039253.

Pierrard, Olivier (2008). 'Commuters, residents and job competition', *Regional Science and Urban Economics*, 38(6), pp. 565–577.

Pires, I. and Nunes, F. (2018). 'Labour mobility in the Euroregion Galicia–Norte de Portugal: constraints faced by cross-border commuters', *European Planning Studies*. Taylor & Francis, 26(2), pp. 376–395. doi: 10.1080/09654313.2017.1404968.

Pontes, T., Magno, G., Vasconcelos, M., Gupta, A., Almeida, J., Kumaraguru, P. and Almeida, V. (2012). 'Beware of what you share: Inferring home location in social networks', *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*. IEEE, pp. 571–578. doi: 10.1109/ICDMW.2012.106.

Poorthuis, A. and Zook, M. (2017). 'Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media', *Journal of Urban Technology*. Taylor & Francis, 24(4), pp. 115–135. doi: 10.1080/10630732.2017.1335153.

Ralph, David (2015). "'Always on the Move, but Going Nowhere Fast': Motivations for "Euro-commuting" between the Republic of Ireland and Other EU States', *Journal of Ethnic and Migration Studies*. Taylor & Francis, 41(2), pp. 176–195. doi: 10.1080/1369183X.2014.910447.

Ramadier, T., Lee-Gosselin, M. and Frenette, A. (2005). 'Conceptual Perspectives for Explaining Spatio-Temporal Behaviour in Urban Areas: Behavioural Foundations', pp. 87–100. doi: 10.1108/9781786359520-004.

Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S. and Waller, S.T. (2017). "Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges Transp. Res. Part C Emerg. Technol., 75 (2017), pp. 197-211, 10.1016/j.trc.2016.12.008"

Schmitz, F., Drevon, G. and Gerber, P. (2012). 'La mobilité des frontaliers du Luxembourg : dynamiques et perspectives', *LES CAHIERS DU CEPS/INSTEAD*.

Sheller, M. and Urry, J. (2006). 'The new mobilities paradigm', *Environment and Planning A*, 38(2), pp. 207–226. doi: 10.1068/a37268.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P. and Rana, O. (2013). 'Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter', *Sociological Research Online*, 18(3), pp. 1–11. doi: 10.5153/sro.3001.

Sloan, L. and Morgan, J. (2015). 'Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter', *PLoS ONE*, 10(11), pp. 1–15. doi: 10.1371/journal.pone.0142209.

Sohn, Christophe (2014). 'Modelling Cross-Border Integration: The Role of Borders as a Resource', *Geopolitics*. Routledge, 19(3), pp. 587–608. doi: 10.1080/14650045.2014.913029.

STATEC (2016). 'Statistiques en bref Statistische Kurzinformationen 2016'. Available at: http://www.grande-region.lu/eportal/pages/ViewDocument.aspx?contentid=b4449d05-4c28-4e1f-ac1b-cf8de199afcb (Accessed: 20 April 2019).

Strüver, Anke (2002). 'Significant insignificance – boundaries in a borderless European Union: Deconstructing the Dutch-German transnational labor market', *Journal of Borderlands Studies*. Routledge, 17(1), pp. 21–36. doi: 10.1080/08865655.2002.9695580.

Tenkanen, Henrikki (2013). *Geographic knowledge discovery from sparse GPS-data - Revealing spatio-temporal patterns of Amazonian river transports*. University of Helsinki.

Tenkanen, Henrikki (2017). *Capturing time in space: Dynamic analysis of accessibility and mobility to support spatial planning with open data and tools*. University of Helsinki.

The Government of the Grand Duchy of Luxembourg (2018). *Luxembourg in the Greater Region — government.lu*. Available at: https://gouvernement.lu/en/dossiers/2018/grande-region.html (Accessed: 5 April 2019).

The Office of the Data Protection Ombudsman (2018). *EU:n tietosuoja-asetus - usein kysyttyjä kysymyksiä - Tietosuojavaltuutetun toimisto*. Available at: https://tietosuoja.fi/gdpr (Accessed: 13 April 2019).

Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C. M., Blondel, V., Smoreda, Z., González, M. C. and Colizza, V. (2014). 'On the Use of Human Mobility

Proxies for Modeling Epidemics', *PLoS Computational Biology*, 10(7). doi: 10.1371/journal.pcbi.1003716.

Tobler, Waldo R. (1970) 'A Computer Movie Simulating Urban Growth in the Detroit Region', *Economic Geography*. Clark University, Wiley, 46, pp. 234–240. Available at: http://www.jstor.org/stable/143141.

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H. and Di Minin, E. (2019). 'Social Media Data for Conservation Science: a Methodological Overview', *Biological Conservation*. Elsevier, 233(January), pp. 298–315. doi: 10.1016/j.biocon.2019.01.023.

Twitter (2019). *Get Tweet timelines*. Available at: https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline.html (Accessed: 21 April 2019).

United Nations Statistics Division (2019). *Methodology – Standard country or area codes for statistical use (M49)*. Available at: https://unstats.un.org/unsd/methodology/m49/ (Accessed: 6 September 2019).

Uprichard, Emma (2013). *Focus: Big Data, Little Questions? | Discover Society*, *Discover Society*. Available at: https://discoversociety.org/2013/10/01/focus-big-data-little-questions/ (Accessed: 20 April 2019).

UrbiStat (no date). *Demographic statisticsRegion of NORDRHEIN-WESTFALEN, population density, population, average age, families, foreigners*. Available at: https://ugeo.urbistat.com/AdminStat/en/de/demografia/dati-sintesi/nordrhein-westfalen/5/2 (Accessed: 5 April 2019).

Vanhoof, M., Reis, F., Ploetz, T. and Smoreda, Z. (2018). 'Assessing the quality of home detection from mobile phone data for official statistics'. doi: 10.2478/jos-2018-0046.

Wiesböck, L., Verwiebe, R., Reinprecht, C. and Haindorfer, R. (2016). 'The economic crisis as a driver of cross-border labour mobility? A multi-method perspective on the case of the Central European Region', *Journal of Ethnic and Migration Studies*. Taylor & Francis, 42(10), pp. 1711–1727. doi: 10.1080/1369183X.2016.1162354.

Willberg, Elias (2019). *Bike sharing as part of urban mobility in Helsinki – a user perspective*. University of Helsinki.

Wojcik, S., Messing, S., Smith, A., Rainie, L. and Hitlin, P. (2018). 'Bots in the Twittersphere Pew Research Center, April 2018', (April). Available at: http://assets.pewresearch.org/wp-content/uploads/sites/14/2018/04/06160833/PI_2018.04.09_Twitter-

Bots_FINAL.pdf%0Ahttp://www.pewinternet.org/2018/04/09/bots-in-the-twittersphere/ (Accessed: 5 April 2019).

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Peña Gangadharan, S., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A. and Pasquale, F. (2017). 'Ten simple rules for responsible big data research', *PLoS Computational Biology*, 13(3), pp. 1–10. doi: 10.1371/journal.pcbi.1005399.